НАЦИОНАЛЬНАЯ АКАДЕМИЯ НАУК УКРАИНЫ
ИНСТИТУТ ПРОБЛЕМ РЕГИСТРАЦИИ ИНФОРМАЦИИ
НАН УКРАИНЫ

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И БЕЗОПАСНОСТЬ

МАТЕРИАЛЫ XVIII МЕЖДУНАРОДНОЙ НАУЧНО-ПРАКТИЧЕСКОЙ КОНФЕРЕНЦИИ

ВЫПУСК 18

Киев – 2018

Рекомендовано к печати Ученым советом
Института проблем регистрации информации НАН Украины
(протокол № 16 от 18 декабря 2018 г.)

Информационные технологии и безопасность. Материалы XVIII Международной научно-практической конференции ИТБ-2018. – К.: ООО "Инжиниринг", 2018. – 359 с. ISBN 978-966-2344-69-1.

В сборник вошли материалы докладов, представленных на XVIII Международной научно-практической конференции «Информационные технологии и безопасность» (ИТБ-2018, 27 ноября 2018 года, г. Киев, Украина).

В сборнике представлены статьи, посвященные вопросам кибернетической безопасности критических инфраструктур, технологиям информационно-аналитических исследований на основе открытых источников информации, онтологическому подходу, семантическим сетям, сценарному анализу при обеспечении информационной поддержки принятия решений, компьютерному моделированию процессов и систем, актуальным проблемам технологического и правового обеспечения информационной и кибернетической безопасности.

Для специалистов в области информационных технологий, информационной безопасности, информационного права а также для аспирантов и студентов старших курсов высшей школы соответствующих специальностей.

ISBN 978-966-2344-69-1

ПІДВИЩЕННЯ ЖИВУЧОСТІ АВТОМАТИЗОВАНИХ СИСТЕМ ОРГАНІЗАЦІЙНОГО УПРАВЛІННЯ ЯК ШЛЯХ ДО БЕЗПЕКИ КРИТИЧНИХ ІНФРАСТРУКТУР

О.Г. Додонов, О.С. Горбачик, М.Г. Кузнєцова
Інститут проблем реєстрації інформації Національної академії наук України, м. Київ, Україна
dodonov@ipri.kiev.ua, ges@ipri.kiev.ua, margle@ipri.kiev.ua

Анотація. *Розглянуто проблеми організації безпечного функціонування критичних інфраструктур. Проаналізовано загрози, підходи до аналізу і оцінювання ризиків безпеки критичних інфраструктур, світовий досвід із захисту об'єктів критичних інфраструктур (ОКІ). Показано, що складність оцінки безпеки критичної інфраструктури пов'язана з різнорідністю її складових, емержентністю інфраструктури, неповнотою знань про можливі відмови і ризики, невизначеністю даних про стан та функціонування ОКІ і власне інфраструктури. Визначено основні особливості автоматизованих систем організаційного управління (СОУ) критичних інфраструктур і функціїї СОУ у забезпеченні сталості і безпеки функціонування критичної інфраструктури у разі виникнення зовнішніх чи внутрішніх загроз на ОКІ, а також при порушеннях в процесах управління самих автоматизованих СОУ. Підвищення живучості автоматизованої СОУ, зокрема її технічної складової, дозволяє забезпечити управління об'єктами критичних інфраструктур таким чином, щоб своєчасно відреагувати на загрози критичній інфраструктурі, запобігти переходу інфраструктури або її складових у небезпечний стан, виконати напрацювання відповідних управлінських рішень. Запропоновано оцінку живучості автоматизованої СОУ як здатності системи до реалізації певного комплексу задач і досягнення системної цілі функціонування.*

Ключові слова: *безпека критичних інфраструктур, автоматизовані системи організаційного управління, живучість.*

Вступ

Термін «критична інфраструктура» не має усталеного тлумачення, і в документах кожної країни, як правило, привноситься в нього свій зміст і своя специфіка, та зазвичай як критичні (критично важливі, ключові) визначаються інфраструктури, від яких залежить суспільний порядок, економічна стабільність і національна

безпека. Такі інфраструктури забезпечують життєво важливі потреби суспільства і визначають рівень його розвитку й благополуччя [1]. Будь-яка критична інфраструктура являє собою велику складну систему стратегічного масштабу, і є сукупністю значної кількості елементів різного типу, поєднаних зв'язками різної природи і маючих загальну властивість, яка відрізняється від властивостей окремих елементів. Функціонування критичних інфраструктур забезпечує підтримування життєво важливих функцій в суспільстві, захист базових потреб його членів і формування у них відчуття безпеки і захищеності.

Системи організаційного управління (СОУ) об'єктами критичної інфраструктури (ОКІ) і всією інфраструктурою мають забезпечувати їх нормальне функціонування для отримання бажаного кінцевого результату у визначених умовах функціонування і адекватне реагування на інциденти безпеки, що виникають на ОКІ. СОУ ОКІ являють собою складні ієрархічні соціотехнічні системи з великою кількістю різнорідних елементів, що мають цілеспрямовану поведінку і постійно взаємодіють із мінливим зовнішнім середовищем. Ці системи повинні гарантувати сталість і безпеку функціонування критичної інфраструктури у разі виникнення зовнішніх чи внутрішніх загроз на ОКІ, а також при порушеннях в процесах управління власне СОУ.

Засоби СОУ мають забезпечити своєчасне розпізнавання загрози і моменту настання критичної (надзвичайної, нештатної) ситуації, обрати адекватний рівень опрацювання, ініціювати процеси протидії, компенсації чи адаптації до ситуації, створити умови продовження функціонування критичної інфраструктури у повному або частковому обсязі, у разі необхідності активувати процедури поступової деградації чи безпечної зупинки функціонування.

Існуюча тенденція до ускладнення ОКІ як технологічного, так і експлуатаційного характеру призводить до зростання кількості елементів СОУ, задіяних у процесах моніторингу і управління, урізноманітнення структур взаємодії в СОУ, а це породжує збільшення кількості і різноманітності видів і типів ризиків, які можуть викликати порушення у функціонуванні ОКІ і в інфраструктурі в цілому. Завдання виявлення ключових об'єктів (або їх сукупності), небажаний спрямований вплив на які може найбільше зашкодити функціонуванню всієї інфраструктури і призвести до порушення життєво важливих процесів, оцінювання наслідків негативних впливів і розробка механізмів зниження ризиків для критичної інфраструктури належать сьогодні до

пріоритетних. Одним зі шляхів підвищення безпеки функціонування ОКІ і критичних інфраструктур в цілому може стати живучість СОУ ОКІ, зокрема технічної складової СОУ.

Питання безпеки критичних інфраструктур України

Розвинуті країни світу вже добре усвідомлюють існуючі загрози для критичних інфраструктур. У США як частина національного дивізіону кібербезпеки (NCSD) функціонує спеціальна програма захисту систем управління і працює спеціальна команда реагування на кіберзагрози у промислових системах (ICS-CERT - Industrial Control Systems Cyber Emergency Response Team). Європейською комісією розроблено глобальну стратегію захисту критичної інфраструктури («The European Programme for Critical Infrastructure Protection»), яка передбачає комплекс заходів з профілактики, запобігання і реагування на терористичні атаки в Європі.

Для критичних інфраструктур України загрозливими факторами є бойові дії на українській території, висока зношеність основних фондів, серйозні проблеми із забезпеченням екологічної та техногенної безпеки, загрози виникнення аварій на об'єктах підвищеної небезпеки: шахтах, об'єктах електроенергетики, хімічних і металургійних підприємствах і мережах життєзабезпечення, як внаслідок їх випадкового пошкодження або втрати контролю над технологічними процесами, так і в результаті терористичних актів і диверсій.

У 2015 і 2016 роках в Україні мали місце порушення у функціонуванні критичних інфраструктур. Так, через деструктивні дії зловмисників у деяких регіонах країни припинялось постачання електроенергії для тисяч споживачів, виходили з ладу електронні системи «Укрзалізниці», безпосередньо перед плануванням соціальних виплат і пенсій на межі знищення були дані Держказначейства [2]. Міжнародна компанія у сфері безпеки CyberX виявила сліди проведення широкомасштабної операції по кібершпіонажу (операція BugDrop) в Україні. CyberX встановив, що у рамках операції BugDrop мішенню були також об'єкти критичних інфраструктур. Особливості операції не дозволили CyberX стверджувати, що атака спонсорувалась якоюсь країною чи певною групою хакерів.

Сьогодні захист критичної інфраструктури та підвищення рівня її стійкості визнані як пріоритетні у сфері безпеки України. Основним підходом прийнятий *all hazards approach* - забезпечення захисту від

усіх видів загроз. Для України визначають наступні основні категорії загроз критичній інфраструктурі [1]:

1) аварії та технічні збої, зокрема, авіаційні катастрофи, ядерні аварії, пожежі, аварії у системах енергозабезпечення, викиди небезпечних речовин, відмови систем, аварії та надзвичайні події, обумовлені недбалістю, організаційними помилками тощо;

2) небезпечні природні явища, зокрема, надзвичайні погодні умови, лісові, степові та торф'яні пожежі, сейсмічні явища, епідемії та пандемії, космічні явища, урагани, торнадо, землетруси, цунамі, повені і т. ін.;

3) зловмисні дії, зокрема, зловмисні дії груп або окремих осіб, таких як терористи, злочинці і диверсанти, а також військові дії в умовах війни.

До особливо небезпечних належать комбіновані загрози та загрози, реалізація яких може призвести до катастрофічних і різноманітних каскадних ефектів внаслідок взаємозалежності елементів критичної інфраструктури.

Безпека критичних інфраструктур і функції систем організаційного управління

Складність створення моделі оцінки безпеки критичної інфраструктури в цілому пов'язана з різнорідністю її складових, емержентністю інфраструктури, неповнотою знань про можливі відмови і ризики, отриманням даних про стан та функціонування ОКІ і власне інфраструктури у різних кваліметричних шкалах.

Найчастіше для оцінки ризиків для критичних систем застосовують методи детерміністського аналізу DSA (Deterministic Safety Assesment) та імовірнісного аналізу PSA (Probabilistic Safety Analysis) [3].

Детерміністський аналіз DSA передбачає послідовний аналіз поведінки інфраструктури на множині правдоподібних сценаріїв розвитку аварій з використанням визначених правил і гіпотез стосовно стану підсистем, їх характеристик, дій оператора і т.д. з обмеженням щодо технічної можливості настання певної аварії. Нажаль цей підхід не дозволяє врахувати всі існуючі невизначеності.

Імовірнісний аналіз PSA використовується для оцінки імовірності великих аварій, зокрема на атомних електростанціях. Основою імовірнісного підходу є системний аналіз можливих сценаріїв, а також послідовне дослідження аварій, включаючи вихідні події, шляхи розвитку аварійних ситуацій з урахуванням накладення

відмов систем. Спершу визначаються послідовності подій, які можуть призвести до аварії, а потім виконується оцінювання стану критичного об'єкта (наприклад, цілісність реактора) і можливе розповсюдження наслідків аварії (радіоактивні викиди в атмосферу). На останньому етапі здійснюється оцінка впливу аварії (радіонуклідів) на здоров'я людей. Оскільки великі аварії є досить рідкісними подіями, статистичних даних недостатньо для застосування класичного імовірнісного підходу.

При застосуванні методу аналізу виду і наслідків критичних відмов FMECA (Failure Modes, Effects and Critical Analysis) систематично, шляхом послідовного розгляду інфраструктурних підсистем, визначаються всі можливі види відмов, пошкоджень, аварійних ситуацій та їх результуючий вплив на інфраструктуру і оточуюче середовище. Суть FMECA полягає у визначенні впливу кожного потенційного дефекту (відмови) на функціональність інфраструктури як системи у цілому, і впорядкування відмов відповідно до величини очікуваного збитку. Цей метод дозволяє провести досить повний якісний аналіз причин і наслідків відмов елементів інфраструктури, але він трудомісткий і не враховує можливу деградацію ОКІ і інфраструктури, час настання і залежність відмов [4].

Іноді для аналізу ризиків застосовують модифікації цих основних методів, частково долаючи зазначені недоліки. Та сьогодні все частіше починають залучати для аналізу методи штучного інтелекту, лінгвістичні методи, нечіткі моделі для уточнення основних видів невизначеностей, урахування залежності подій, зміни критичності відмов при наявності залежності відмов.

Методи нечіткої математики, такі як нейроні мережі, нечітка логіка, генетичні алгоритми вбудовуються у технології нового покоління – «м'яких обчислень» (soft computing), які використовуються для управління складними системами з дефіцитом апріорної інформації в умовах невизначеності і дозволяють застосувати досвід експертів, їх знання для ризик-аналізу у процесі управління [5].

Складність і трудомісткість методів аналізу ризиків, складність математичних моделей критичних інфраструктур, неможливість врахування широкого спектру факторів, зокрема нових можливостей щодо дистанційного ураження об'єктів критичної інфраструктури, реалізації загроз виникнення аварій змушують шукати нові підходи.

Враховуючи, що безпека критичних інфраструктур України та їх захист, згідно [1], мають забезпечуватись комплексом заходів,

реалізованих у нормативно-правових, організаційних, технологічних інструментах, спрямованих на забезпечення як фізичної (фізичний захист), експлуатаційної, так й операційної безпеки та стійкості критичної інфраструктури, доцільно спрямувати зусилля на підвищенні якості СОУ ОКІ.

Автоматизовані системи організаційного управління є складовими будь-якої сучасної критичної інфраструктури [6]. Вони являють собою комплекси апаратних і програмних засобів, інформаційних систем і інформаційно-телекомунікаційних мереж, призначені для вирішення задач оперативного управління і контролю за різними процесами та технічними об'єктами в рамках організації виробництва або технологічного процесу об'єкта критичної інфраструктури і інфраструктури в цілому.

Аналіз сучасних тенденцій розвитку автоматизованих систем управління дозволяє припустити, що буде зростати доля важливих для безпеки критичних інфраструктур функцій, які будуть реалізовуватись на базі комп'ютерних систем. Під безпекою критичної інфраструктури у такому випадку розуміють незалежність від неприйнятного ризику [7], тобто забезпечення такого стану інфраструктури, у якому ризик нанесення шкоди людині, суспільству, країні скорочується до прийнятного рівня.

На СОУ, як складову критичної інфраструктури, у рамках забезпечення безпеки ОКІ і власне інфраструктури покладається виконання наступних функцій:

- реалізація процесів організаційного управління таким чином, щоб не допустити перехід інфраструктури або її складових у потенційно небезпечний стан;
- відключення відповідного технічного об'єкта при появі або реалізації загрози переходу у небезпечний (аварійний) стан;
- прогнозування, оцінювання і мінімізація ризиків безпеці в ході функціонування об'єкта, напрацювання відповідних управлінських рішень.

Протягом всього життєвого циклу інфраструктури у рамках СОУ має виконуватись постійний аналіз процесів функціонування, моніторинг стану, оцінка ризику виникнення загроз, прогнозування наслідків реалізації загроз, розробка стратегій забезпечення безпеки[8]. Під час небажаних впливів засобами СОУ виконується інтерпретація даних щодо стану інфраструктури і окремих її об'єктів, діагностика з метою виявлення і розпізнавання загроз безпеці, моніторинг для виявлення у реальному масштабі часу відхилення тих чи інших параметрів функціонування, прогнозування

наслідків певних подій або явищ, планування дій для працездатних об'єктів, здатних виконати в повному обсязі чи частково певні функції. У той же час СОУ має підтримувати один із визначених для неї режимів функціонування, проводити контроль ходу і результати виконання управлінських рішень, забезпечувати напрацювання й прийняття рішень виконанням відповідних процедур, необхідною інформацією і ресурсами.

Оцінка живучості СОУ ОКІ у рамках проблеми безпеки функціонування критичних інфраструктур

Від рішень, що напрацьовуються в СОУ, суттєво залежить безпека функціонування об'єктів критичних інфраструктур, особливо у разі розвитку аварійної ситуації, тобто в умовах, коли відсутня можливість чіткого передбачення результатів управляючих впливів. Функціональна стабільність СОУ у такому випадку стає фактором і умовою безпеки об'єктів критичних інфраструктур. Показником функціональної стабільності СОУ може слугувати оцінка живучості, що характеризує можливості системи до збереження своєї функціональності у постійно змінних умовах внутрішнього і зовнішнього середовища.

СОУ ОКІ належать до класу соціотехнічних систем і являють собою складні системні утворення, складовими яких є техніко-технологічні підсистеми, відповідні системи діяльності (системи ролей і функцій обслуговуючого і управлінського персоналу) та зовнішнє середовище, активно взаємодіюче із підсистемами. Знання про соціотехнічні системи завжди принципово не повні і не можуть бути повними, оскільки важко повністю визначити структуру зв'язків і відношень, що виникають при функціонуванні систем. Функціонування СОУ відбувається в умовах невизначеності факторів впливу, постійної мінливості середовища функціонування, неможливості чіткого врахування його реакції на дії системи і відповіді системи на зовнішні впливи, тобто в умовах прояву таких фундаментальних системних властивостей, як живучість. Саме завдяки притаманній їй живучості система може зберігатись як ціле у непередбачуваних, іноді екстремальних, умовах, пристосовуватися до них, змінюючи поведінку, структуру чи загальносистемну ціль функціонування [9].

Якісні оцінки та кількісні показники живучості СОУ ОКІ є інтегральними характеристиками системи.

У відповідності із загальною теорією систем будь-яку систему $\mathfrak{I}$, зокрема і СОУ ОКІ, можна визначити наступним чином: $\mathfrak{I} =$

$\langle G, \Re, \Phi \rangle$, де G – множина елементів системи; $\Re$ – система чинних правил, за якими функціонує система; Φ – процес функціонування, визначений на множині G згідно комплексу правил $\Re$; він може бути поданий як $G \xrightarrow{R(*)} \Phi$.

У загальному випадку живучість системи залежить від множини параметрів, що характеризують систему, задач, які вирішуються нею, зовнішнього середовища та типу, ступеню і динаміки їх взаємодії. Якщо система $\Im$ у «стані живучості», то це означає, що системою досягається ціль функціонування, тобто виконується комплекс задач $\phi = (\phi_1, \phi_2, ..., \phi_n)$ із заданою якістю та необхідною ефективністю. Небажані впливи повинні компенсуватися наявними у системі механізмами підтримки живучості, що передбачає розв'язання, зокрема задач моніторингу, ідентифікації, діагностики, відновлення тощо. «Стан живучості» характеризується стабільністю і передбачуваністю (очікуваністю результатів) функціонування системи $\Im$, тобто СОУ ОКІ виконують усі управлінські функції.

Оцінкою живучості системи може слугувати функціонал, заданий на деякій множині параметрів, які впливають на стан системи $\Im$, а саме:

$$\Psi = f(S, B, |S|, \Delta T, U, Q, W, \Lambda),$$

S – структура системи $\Im$; B – поведінка системи; $|S|$ – «стан здатності» системи; ΔT – часова надмірність; U – управління; Q – вектор допустимої якості виконання функцій; W – множина станів, у які може перейти система $\Im$ через впливи зовнішнього середовища; Λ – множина параметрів, що визначають характер, ступінь, топологію і динаміку впливу зовнішнього середовища на систему $\Im$.

Якщо система $\Im$ переходить у стан, коли забезпечується рішення деякого комплексу задач $\phi = (\phi_1, \phi_2, ..., \phi_n)$, то це означає здатність системи $\Im$ реалізувати будь-яку задачу ϕ_i з комплексу задач $\phi = (\phi_1, \phi_2, ..., \phi_n)$ у будь-якому зі стані $w_j \in W$, зокрема, і у станах $w_j \in (w_1, w_2, ..., w_d)$, що характеризуються як відмовами технічних

засобів (відмовами у техніко-технологічній складовій), так і «хибними діями» – порушеннями (навмисними/ ненавмисними) чи помилками обслуговуючого або управлінського персоналу.

Принципово розрізняють три типи $|S| = \{|S|_t\}, t = 1,2,3$, у які може переходити система $\Im$, а саме [7]:

– стан $|S|_1$, при якому у системі забезпечується вирішення усього комплексу задач $\phi = (\phi_1, \phi_2, ..., \phi_n)$ із заданою якістю та необхідною ефективністю у будь-якому із станів $w_j \in W$;

– стан $|S|_2$, при якому у системі забезпечується вирішення лише деякої підмножини $\phi^* \subset \phi$ у будь-якому із станів $w_j \in W$;

– стан $|S|_3$, при якому у системі забезпечується вирішення лише якоїсь однієї із задач комплексу $\phi = (\phi_1, \phi_2, ..., \phi_n)$ у будь-якому із станів $w_j \in W$.

Для СОУ ОКІ перехід стани типів $|S|_2$ та $|S|_3$ означає, що в СОУ ОКІ мають місце порушення у роботі технічних засобів або «хибні дії» з боку персоналу, тому й відбувається звуження множини задач, що реалізуються системою.

У загальному випадку здатність системи до вирішення комплексу задач $\phi = (\phi_1, \phi_2, ..., \phi_n)$ і відповідно досягнення цілі функціонування можна достатньо повно характеризувати наступною матрицею

$$M(|S|) = \left\| m_{ij}(|S|) \right\|, \quad i = \overline{1,n}, \quad j = \overline{1,d} \quad ,$$

$$m_{ij}(|S|) = \begin{cases} 1, & \textit{якщо у системі, що знаходиться у стані } w_j, \\ \textit{існує можливість виконання задачі } \varphi_i \textit{ з необхідною якістю} \\ 0, \textit{ в іншому випадку} \end{cases}$$

Якщо задати матрицю $M(|S|)$ і розподіл ймовірностей знаходження системи $\Im$ у будь-якому зі станів $w_j \in (w_1, w_2, ..., w_d)$, то живучість системи $\Im$ буде визначено, і тоді для системи $\Im$ у

якості оцінки її живучості замість функціонала Ψ можна використати простіший, записаний у матричній формі:

$$\Psi = V \times M(|S|) \times P, \qquad (1)$$

де $V = \|v_1, v_2, ..., v_n\|$ – вектор коефіцієнтів важливості задач з множини ϕ, що реалізуються системою $\Im$, $P = \|p_1, p_{2,...,} p_d\|$ – вектор імовірності стану w_j. Коефіцієнт важливості задачі має характеризувати втрати у функціональності критичної інфраструктури (відносні) у випадку невиконання СОУ ОКІ цієї задачі. При оцінці живучості СОУ ОКІ за (1) найскладнішим є формування матриці $M(|S|)$, елементи якої є булеві функції, що задані на множині параметрів, які впливають на стан системи. Слід зазначити, що з усієї множини можливих станів СОУ ОКІ при оцінці живучості доцільно розглядати лише ті, у яких погіршуються показники якості функціонування системи, наприклад, час реалізації управлінських функцій, чи імовірність вирішення деякої задачі наближається до нуля.

Джерела

1. Зелена книга з питань захисту критичної інфраструктури в Україні. Київ, 2015. 35 с.
2. *Горбачик О.* Проблеми і задачі забезпечення безпеки функціонування об'єктів критичних інфраструктур. Реєстрація, зберігання і обробка даних: зб. наук. праць за матеріалами Щорічної підсумкової наукової конференції 17-18 травня 2017 року, ІПРІ НАН України. Київ, 2017. С. 106-109.
3. Безопасность критических инфраструктур: математические и инженерные методы анализа и обеспечения/ Под ред. Харченко В.С. – МОН Украины, Национальный аэрокосмический университет им. Н.Е.Жуковского «ХАИ», 2011. – 641 с.
4. Failure Mode, Effects & Criticality Analysis (FMECA). URL: https://quality-one.com/fmeca/
5. *Dogan Ibrahim.* An Overview of Soft Computing. URL: https://www.sciencedirect.com/science/article/pii/S1877050916325467

6. *Додонов О.Г.* Комп'ютерне моделювання процесів організаційного управління, Вісник НАН України, 2016, № 1. С. 69 – 77.
7. *Basilio A., Landrini F., Novelli G., Landrini G., Baldrighi M.* Functional Safety of Safety-Related Systems. Manual for Plant Engineering and Maintenance. Italy, G.M. International S.r.l, Villasanta, 2008. 388 p.
8. *Кузнєцова М.Г.* Системи організаційного управління та безпека об'єктів критичних інфраструктур. Реєстрація, зберігання і обробка даних: зб. наук. праць за матеріалами Щорічної підсумкової наукової конференції 17-18 травня 2017 року, ІПРІ НАН України. Київ, 2017. С. 109-111.
9. *Додонов О.Г., Кузнєцова М.Г., Горбачик О.С.* Методологічні аспекти створення корпоративних інформаційно-аналітичних систем підвищеної живучості // Реєстрація, зберігання і обробка даних, 2012. – N 3, Т.14. – С. 58-69.

НАЦІОНАЛЬНА ПРОГРАМА ІНФОРМАТИЗАЦІЇ ЯК ІНСТРУМЕНТ СТРАТЕГІЧНОГО УПРАВЛІННЯ УКРАЇНИ

І.Б. Жиляєв, А.І. Семенченко, В.М. Фурашев

Особливостями сучасного суспільно-політичного та соціально-економічного розвитку є його значний динамізм, багатовекторність, невизначеність, глобальність та суперечливість, що значно ускладнює процеси управління цим розвитком. Традиційні підходи та методи публічного управління та адміністрування виявилися недостатньо спроможними ефективно розв'язувати значну кількість сучасних проблем, які стає все важче прогнозувати. Однак, збільшення кількості, масштабів та рівня загроз та ризиків для громадян, суспільства та держави актуалізує необхідність посилення впливу інститутів громадянського суспільства на формування та реалізацію публічної політики, яка легітимізується у відповідних стратегічних актах: зокрема, в Стратегії соціально-економічного розвитку Європейського Союзу на період до 2020 року «Європа 2020», а також в українських: Стратегії сталого розвитку «Україна – 2020», Стратегії реформування державного управління України на 2016-2020 роки, Програмі діяльності Кабінету Міністрів України, Концепції розвитку цифрової економіки та суспільства України на 2018-2020 тощо.

Особливе місце в формуванні парадигми соціально-економічного розвитку займає широке та стрімке впровадження інформаційно-комунікаційних технологій (ІКТ) у всіх сферах. Це надає новий зміст процесам глобалізації, якому притаманно ускладення та непедбачуваність економічного та суспільно-політичного розвитку послідовному оновленню концепцій розвитку із застосуванням ІКТ: інформаційного суспільства та суспільства знань, цифровізації, а також трансформація 3-го технологічного укладу в 4-й (4-а технологічна революція), актуалізація проблем інформаційної безпеки та кібербезпеки тощо.

Національна програма інформатизації (НПІ) визначає стратегію розв'язання проблеми забезпечення інформаційних потреб та інформаційної підтримки соціально-економічної, екологічної, науково-технічної, оборонної, національно-культурної та іншої діяльності у сферах загальнодержавного значення і тому обґрунтовано відноситься до стратегічних документів. Її механізм формування та виконання було запроваджено 20 років тому у

вигляді Законів України «Про Концепцію національної програми інформатизації» та «Про національну програму інформатизації», «Про завдання національної програми інформатизації на 1998-2000 роки» і низки підзаконних актів. Враховуючі багаторічний досвід реалізації НПІ, актуальною є проблема аналізу та оцінки її організаційно-правових механізмів публічного управління, визначення основних факторів впливу на НПІ та формування пропозицій щодо їх удосконалення.

Залежно від ступеня досяжності кінцевих та проміжних цілей в інформатизації виокремлюються три етапи її розвитку:

1) створення політичних, організаційних, законодавчих, соціальних, економічних та технічних умов формування та початкового задоволення інформаційних потреб громадян, суспільства та держави;

2) розвиток інформаційної інфраструктури та забезпечення умов для її включення у світову;

3) повне та якісне забезпечення інформаційних потреб громадян, суспільства, держави та бізнесу .

Але з точки зору кількості та якості отриманих за 20 років позитивних результатів НПІ не можна вважати успішною. Так, майже не були досягнуті цілі НПІ, що були визначені як в її концептуальній частині, так і в переліках завдань. Деякі з її складових, а саме програми та проекти інформатизації органів місцевого самоврядування так і не були впроваджені, значна кількість регіональних проектів та завдань з інформатизації виконувались поза межами НПІ.

Останні роки визначились негативні тенденції в механізмах формування та реалізації НПІ:

зменшення її координуючого впливу і, як наслідок, зменшення її організаційного, кадрового та фінансового забезпечення, всупереч зростаючої динаміки впровадження ІКТ в усі сфери життєдіяльності суспільства, людини та держави;

зменшення впливу громадськості, бізнесу та науки на формування та виконання НПІ (виключення консультативно-дорадчих органів НПІ, відміна щорічного звітування про стан розвитку інформаційного суспільства та інформатизації.

Серед основних недоліків існуючої НПІ також відзначаються наступні:

спрямованість в основному на розвиток інфраструктури органів публічної влади з малочисленими локальними дослідженнями та розробками;

негнучка, надмірно забюрократизована структура, не узгоджена з іншими інструментами публічного управління;

не орієнтованість на кінцевих споживачів публічних послуг;

значна затримка затвердження щорічних завдань (проектів) та виділення бюджетних коштів;

«інерційність» у запровадження нових глобальних ініціатив: розвитку цифрової економіки, електронного та відкритого уряду, електронної демократії, електронної комерції тощо, а також механізмів державно-приватного та державно-громадського партнерства;

фактична відсутність необхідного наукового, інформаційно-аналітичного, організаційного та фінансового забезпечення, останнє робить НПІ декларативною тощо.

При прийнятті Урядом рішення щодо майбутнього оновлення законодавчого забезпечення НПІ було враховано такі її позитивні якості:

20 річний досвід застосування та необмеженість у часі;

наявність сформованої законодавчої терміносистеми, що достатньо динамічно оновлюється;

достатньо конкретний, детальній, прозорий та відкритий механізм формування та виконання НПІ, якій у першому десятиріччі передбачав активну участь громадськості, експертів та бізнесу у формуванні публічної політики у сфері інформатизації та її реалізації;

законодавчо продуману організаційну структуру управління НПІ, що включає: Генерального державного замовника, якій підпорядкований безпосередньо голові Уряду, державних замовників завдань (проектів), керівника НПІ та керівників галузевих (регіональних) програм(проектів) інформатизації;

створену мережу підрозділів в органах влади, що відповідають за інформатизацію;

опрацьовані механізми управління сукупністю державних, галузевих та регіональних програм та проектів інформатизації; розвинуту, системну та ієрархічну нормативно-правову базу НПІ;

успішність реалізації деяких регіональних програм інформатизації, наприклад, Волинської та Дніпропетровської, та відпрацювання механізму їх взаємодії з програмами інформатизації органів місцевого самоврядування.

В той же час однією з головних організаційних проблем НПІ залишається не координованість дій органів публічної влади в цій сфері, коли створюються державні органи з дублюючими

завданнями та функціями, розробляються та приймаються неузгоджені між собою програмні та планові документи, порушуються зв'язки та управлінська ієрархія суб'єктів управління, що призводить до конкуренції органів влади між собою за бюджетні кошти .

Досвід реалізації НПІ демонструє, що у багатьох випадках органи публічного управління, «наштовхуючись» на складнощі з інформатизацією певної галузі економіки, обмеженістю ресурсного забезпечення тощо прагнуть вирішити проблеми галузі автономно, запровадивши локальні механізми. На це також впливає критичне відношення до самої НПІ, прагнення оцінити її результативність на короткому проміжку часу (прив'язаному до фінансового року). Іншим фактором недооцінки необхідності цілісного стратегічного управління сферою ІКТ є «тиск» лобіювання новітніх міжнародних концепцій, які часто динамічно змінюють одне одну.

Незважаючи на це на сьогодні НПІ залишається єдиним «системним інтегратором», який став своєрідним «стрижнем» інформатизації, яка включає певну кількість організаційно-правових інструментів публічного управління національним розвитком із застосуванням ІКТ, зафіксованих у стратегічних документах щодо: самої Національної програми інформатизації (з 1998); державних цільових програм щодо галузевої комп'ютерізації, запровадження ІКТ, формування комунікаційних систем та мереж тощо (з 2004); розвитку інформаційного суспільства (з 2007); розвитку електронного урядування (з 2010-2015, з 2017); розвитку електронної демократії (з 2017); міжнародної ініціативи «Відкритий Уряд» (з 2012); Стратегії кібербезпеки та Доктрини інформаційної безпеки (з 2016 та 2017 відповідно); Концепції розвитку цифрової економіки та суспільства України на 2018- 2020 роки тощо. Фактично, в Україні демонструвалася певна зміна державних концепцій впровадження ІКТ. Можна виокремити певні етапи змін концепцій розвитку української ІКТ-сфери: 1.0 «Інформатизація»; 2.0 «Інформаційне суспільство»; 3.0 «Е-урядування та е-демократія»; 4.0 «Цифрова епоха (цифровізація економіки, суспільства, влади, країни, міст тощо), які було зафіксовано у відповідних стратегічних нормативно-правових актах.

В цих стратегічних документах ставляться відмінні цілі та завдання, суб'єктами управління, як правило, виступають різні державні органи, а об'єкти управління мало взаємодіють між собою та лише частково стосуються проблем інформатизації, самі акти часто не виходять за рамки концепцій (доктрин), мають переважно

декларативний характер, не будучі підтримані організаційно та ресурсне.

«Інтеграційний» підхід було визначено за результатами відповідних парламентських слухань, де рекомендовано Уряду: «утворити центральний орган виконавчої влади, що забезпечуватиме формування та/або реалізацію державної політики у сферах ІКТ та зв'язку, розвитку інформаційного суспільства, інформатизації, телекомунікацій, програмування, інформаційної безпеки та кібербезпеки, впровадження технологій електронного урядування, електронного документообігу, електронного підпису тощо та передати зазначеному органу повноваження інших органів виконавчої влади, що стосуються сфери ІКТ та зв'язку, чітко розмежувати повноваження між органами виконавчої влади в зазначених сферах відповідно до законодавства Європейського Союзу» .

Інший підхід передбачає розробку та затвердження замість Закону України «Про Концепцію національної програми інформатизації», так званих, «пріоритетних напрямків Національної програми інформатизації», що включають: стратегічні цілі інформатизації, її основні принципи, напрямки, очікувані наслідки реалізації НПІ та формуються центральним органом виконавчої влади, який реалізує державну політику у сферах інформатизації, електронного урядування, формування і використання національних електронних інформаційних ресурсів, розвитку інформаційного суспільства, визначаються урядовою постановою .

Враховуючі основні тенденції розвитку сфери інформатизації, ІКТ, цифрової економіки, реформ, що відбуваються в Україні, пропонується уточнити головну мету НПІ, визначивши пріоритетні напрямки та стратегію реалізації публічної політики у сфері інформатизації із забезпечення інфраструктурних потреб розвитку інформаційного суспільства та цифрової економіки, формування і використання національних електронних інформаційних ресурсів, впровадження сучасних ІКТ у всіх сферах життєдіяльності громадянина, суспільства та держави.

Висновки:

1. У 1998 році в Україні було створено механізм стратегічного управління розвитком інформатизації, який охопив майже всі сфери політичного, соціально-економічного та культурного життя. Однак внаслідок сукупності негативних факторів, насамперед таких як недостатня координація, розпорошеність та обмеженість ресурсів,

НПІ за 20 років лише частково реалізувала планові завдання, перетворилася значною мірою в декларативну, мало демократичну, негнучку бюрократичну структуру. В той же час, незважаючи на всі ці недоліки, НПІ залишається єдиним реально діючим інструментом стратегічного управління розвитком на основі ІКТ, яка в останні роки певною мірою підвищила результативність.

2. Актуальною залишається проблема збереження генеральної спрямованості розвитку країни на прискорене запровадження ІКТ у всі сфери суспільного та економічного життя, забезпечення синергії державних рішень щодо: інформаційного суспільства, відкритого уряду, е-урядування та е-демократії, інформаційної та кібербезпеки тощо.

3. Реалії сучасного розвитку, накопичений український та світовий досвід застосування різних інструментів публічного управління запровадженням ІКТ дозволив обґрунтувати надані пропозиції щодо необхідності модернізації організаційно-правового механізму Програми, основні зміни щодо складу її суб'єктів, державних замовників, об'єктів, їх функцій та завдань, спрямованих на усунення існуючих недоліків та на підвищення рівня демократичності, гнучкості, ефективності, результативності та відповідності міжнародних механізмів.

TRANSFORMATION TEXTS INTO COMPLEX NETWORK WITH APPLYING VISIBILITY GRAPHS ALGORITHMS

D.V. Lande[2] , O.O. Dmytrenko[1], A.A. Snarskii[2]
[1] Institute for information recording, NAS Ukraine, Kiev, Ukraine
[2] National Technical University "Igor Sikorsky Kyiv Polytechnic Institute", Kiev, Ukraine
dwlande@gmail.com, dmytrenko.o@gmail.com, asnarskii@gmail.com

In this article the algorithms of visibility for transforming texts into a complex network is proposed. Key words and concepts from the set of documents which describe some subject domain are extracted. Numeric values are assigned to each word or phrase using GTF-IDF metric, which was proposed in this article instead ordinary TF-IDF metric, that is intended to reflect how important a word is to a document in a collection or corpus. As the result a time series are constructed. A tool in time series analysis – the visibility graph algorithm is used for constructing the graph of subject domain. In this article two actual subject domains ("Information extraction" and "Complex network") are considered for example. The corpora of documents, which are related with actual subject domains, were considered from an open access repository of electronic preprints – arXiv (https://arxiv.org). The proposed algorithm is used for the set of documents, which are related with "Information extraction" and "Complex network". This article shows that applying only GTF metric is more expedient compared with GTF-IDF metric in case when the set of documents describe one subject domain. Also the results of applying the visibility graph algorithm and the compactified horizontal visibility graph algorithm are compared. This article shows, that in some case using the compactified horizontal visibility graph algorithm gives a network of words with more quantity of connections between concepts compared with using the visibility graph algorithm. An open-source visualization and exploration software for all kinds of graphs and networks Gephi and an original package of specially developed Python modules are used for simulation and visualization as an additional tool. The proposed algorithm can be used for visualization some subject domain, and also for information support systems, enabling to reveal key components of the subject domain. Also the results of this article can be used for building UI of information retrieval systems, enabling to make a process of search a relevant information easier.

Keywords: *Set of Documents, Subject Domain, Time Series, Network of Words, TF-IDF, Visibility Graph, Compactified Horizontal Visibility Graph.*

1 Introduction

The development of the Internet caused set of problems, which related, first of all, with the massive quantity of data in the Web-space, including a needless data.

Today in the Internet there is a huge and dynamical information base which is an available for research and analysis. It turned out, that many tasks, which arise during working with the network information space, have much in common with mathematical sciences. This fact opens wide opportunities to applying a powerful mathematical tool [1,2]. Taking into account the problems of the huge dimensionality and the dynamic of information resources in global networks, the knowledge based on discrete mathematics (graph theory, networks theory), pattern recognition (classification, clustering), linguistics, digital signal processing, wavelet analysis and fractal analysis are applied.

Due to terabytes of textual data, that are distributed in networks and have been accumulating dynamically, development of new methods and algorithms for analysis these data is necessary. But also the advantages and disadvantages of algorithms that exist for information retrieval must consider.

A modern development of technologies in some case enable to find relevant information. But the problems of further analytical processing of this information, selection of necessary factual data, detection of development trends in selected subject domain, relation between concepts, events, and forecasting remain unresolved. More of these problems are actual challenges of a semantic processing of huge dynamical sets of textual data.

2 Analysis of recent researches and publications

The subject of this study is actual and most commonly found in various articles of foreign and domestic scientists. For example, in the works [3,4] the main accent makes on development new methods and algorithms, which are appointed to analytical processing of huge sets of textual data. In the works [5,6] authors consider a linguistic processing of

natural language texts, as one of the central problem of intellectualization of information technologies.

In particular, in the works [7-10] the visibility graph algorithm is proposed. Also the method to constructing networks based on the visibility graph algorithm is presented in works [11-15].

3 Review of some visibility algorithms

In this work network of connections between terms and concepts, which go into textual data is building. Building networks of words, the nodes of which are elements of the text, enables to reveal key components of the text. At the same time, the task of determining, which of the important structural elements of the text are also informationally important, is actual.

There are several approach to constructing networks from the texts (so-called language networks), and different ways to interpreting nodes and connections. It leads, accordingly, to various kinds of presenting of such networks. Nodes are connected if corresponding words are either adjacent in the text [16, 17], or are in a single sentence [18], or are syntactically [19, 20] or semantically [21, 22] connected.

3.1 Visibility graph algorithm (VG)

In this article a tool in time series analysis – the visibility graph algorithm [7, 23, 24] is used for constructing the network. This algorithm maps a time series into a network.

For example, the derived graph of visibility for the time series {0.125, 0.063, 0.042, 0.104, 00.125, 0.063, 0.042, 0.104, 0.125, 0.063, 0.042, 0.104} is presented in Fig.. In the graph, every node corresponds, in the same order, to series data. The visibility rays between the data define the links connecting nodes in the graph.

There is a connection between nodes if they are in “line of sight” with each other, i.e., if they can be connected by a line that does not cross any other histogram bar. More formally, the visibility criteria is describes as follows: two arbitrary data values (t_a, y_a) and (t_b, y_b) will have visibility, and consequently will become two connected nodes of the associated graph, if any other data (t_c, y_c) placed between them fulfills:

$$y_c < y_b + (y_a - y_b)\frac{t_b - t_c}{t_b - t_a}.$$

Also in the article [7] is shown that the structure of the time series is conserved in the graph topology: periodic series convert into regular graphs, random series into random graphs, and fractal series into scale-free graphs.

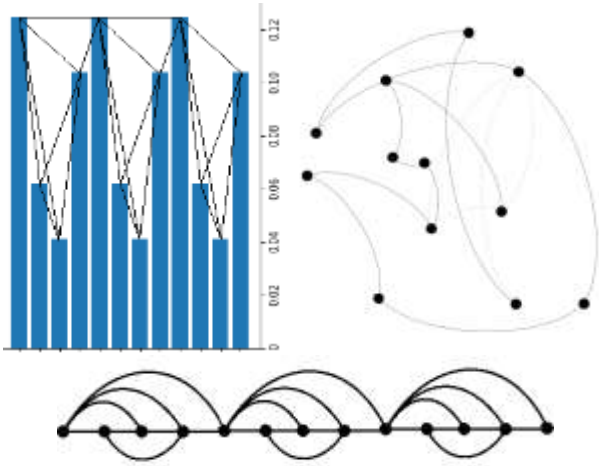


Fig. 1. Example of a time series and the associated graph derived from the visibility algorithm

3.2 Compactified horizontal visibility graph algorithm (CHVG)

In the works [11, 12, 13, 25-27] another algorithm to constructing networks of words – the compactified horizontal visibility graph algorithm (CHVG) is proposed. In general, the process of constructing the language network using the compactified horizontal visibility graph algorithm consists of three stages (Fig. 2). At the first stage the set of nodes, which correspond to the set of words in order of occurrence in the text, are marked on the horizontal axis. At the second stage the horizontal visibility graph is built. Two observations made at times t_i and t_j to be connected in a horizontal visibility graph (HVG) if and only if

$$x_k < \min\{x_i, x_j\}$$

for all t_k with $t_i < t_k < t_j$.

At the third stage, the network, that was obtained at the previous stages, is compactified. As the result, the new network of words – the compactified horizontal visibility graph is obtained.

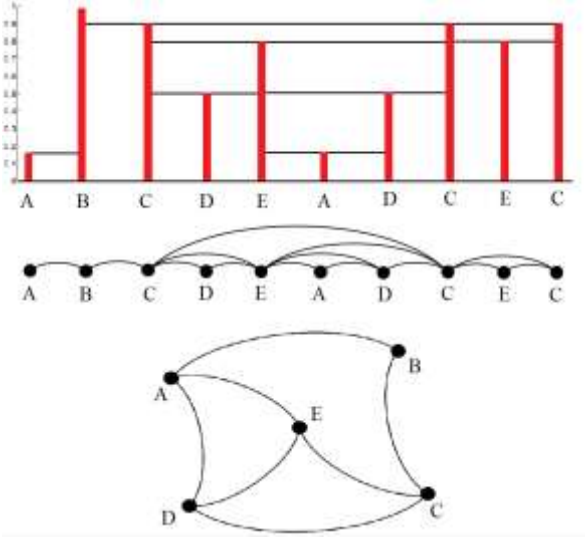


Fig. 2. The stages of building of the compactified horizontal visibility graph (CHVG)

In this manner, the compactified horizontal visibility graph algorithm enables to constructing of network structures based on texts, in which numeric values are assigned in some manner to each word or phrase.

4 Forming of the time series

In this article TF-IDF numeric metric (TF – Term Frequency, IDF — Inverse Document Frequency) is used for forming of the time series. It is an example of function that assigns a number to a word in the text. TF-IDF is the most frequently applied weighting scheme. Also this a numerical statistic is intended to estimating how important a word is to a document in a collection or corpus [28]. The TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. It is often used as a weighting factor in text mining,

information searching and retrieval. Also it can be used as one of the criteria for estimating the relevancy of a document to a search query [29].

TF (term frequency) is a ratio of the number of word occurs in a document to the total number of words in the document. In this manner, the weight of a term (word) t_i that occurs in a document is simply proportional to the term frequency. The term was proposed by Karen Spärck Jones [30],

$$TF = \frac{n_i}{\sum_k n_k},$$

where n_i is the number of occurrences of term (word) i in document; $\sum_k n_k$ is a total number of words in the document.

IDF (inverse document frequency) is an inverse function of the number of documents in which a term occurs. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient. Using IDF reduces the weight of widely used terms (words).

$$IDF = \log \frac{|D|}{|(d_i \supset t_i)|},$$

where $|D|$ is a total number of documents in the corpus; $|(d_i \supset t_i)|$ is the number of documents contain a term t_i ($n_i \neq 0$).

In other words, TF-IDF metric is a product of two members: TF and IDF.

$$TF - IDF = TF \bullet IDF$$

A word has high TF-IDF score in a document if it appears in relatively few documents, but appears in this one, and when it appears in a document it tends to appear many times.

After representation of corpora of documents in a vector view (number of words determines the dimension of the vector), the visibility graph algorithm, which was described above, is used.

4 Presentation of basic material of the research

In this article before using the method to constructing networks from the texts we propose to removing stop word. It enables to removing the words, which are informationally unimportant. We use the stop-dictionary based on various stop-dictionaries, which are available via the

links: https://code.google.com/archive/p/stop-words/downloads/; http://www.textfixer.com/tutorials/common-english-words.php.

Also we use global GTF metric, which looks like

$$GTF = \frac{n_i}{\sum_k n_k},$$

where n_i is the number of occurrences of term (word) i in documents of the corpus; $\sum_k n_k$ is a total number of words in the documents of the corpus.

Words, which are occurred not often within single document, have a low TF metric. But if they occur in every documents of corpora, they at real are informationally important in global context for considered subject domain. That is why in this article we use global TF metric (GTF).

4.1 Example 1

In this article the corpora of 292 documents, which are related with an actual do-main – "Information extraction", were considered from an open access repository of electronic preprints – arXiv (https://arxiv.org) for period of time 2000-2010.

As the result of applying of a proposed method to constructing networks from the texts, the network of key words, which are important structural elements of the subject domain, was obtained (Fig. 3).

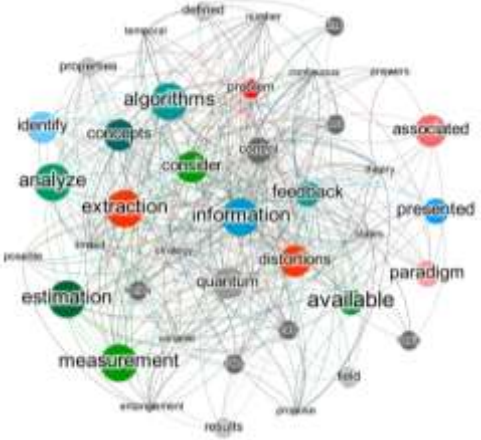


Fig. 3. The network of key words, obtained through the application of the proposed method

Table 1. TOP-40 largest-weight nodes of the network of words constructed from corpora of documents, which describe "Information extraction" subject domain

Weight (GTF-IDF)	Word	Weight (GTF)	Word
0.23	feedback	0.26	quantum
0.23	quantum	0.26	information
0.22	consider	0.25	feedback
0.218	state	0.239	consider
0.214	control	0.237	control
0.211	measurement	0.205	algorithms
0.21	problem	0.203	problem
0.2	states	0.192	measurement
0.18	continuous	0.186	state
0.179	algorithms	0.185	distortions
0.176	estimation	0.185	concepts
0.176	available	0.178	extraction
0.176	distortions	0.171	loss
0.175	concepts	0.168	limited
0.172	theory	0.166	variable
0.165	identify	0.163	estimation
0.163	limited	0.162	analyze
0.162	paradigm	0.159	states
0.155	defined	0.158	properties
0.152	field	0.158	strategy
0.149	questions	0.157	number
0.1489	considered	0.149	available
0.145	properties	0.143	series
0.145	time	0.141	continuous
0.143	content	0.137	theory
0.14	results	0.136	identify
0.135	entanglement	0.113	entanglement
0.132	presented	0.103	temporal
0.128	number	0.101	propose
0.126	rules	0.092	paradigm
0.124	temporal	0.089	field
0.122	propose	0.085	presented
0.101	label	0.084	rules
0.087	discuss	0.082	defined
0.086	process	0.077	associated
0.083	corresponding	0.072	time
0.076	possible	0.071	possible
0.0745	series	0.07	results
0.053	noise	0.067	basis
0.046	evolution	0.055	answers

Based on the results which presented in the Table 1 we can notice that quantity of key words, which are informationally important, is

more in case of applying only GTF metric for the set of documents that describe one subject domain. The key words, such as “information” and “extraction”, which are informationally important for the considered subject domain, are missed in case of using GTF-IDF metric (these key words have a low GTF-IDF). After analyzing the results of research (Table 1) we can make conclusion that applying only GTF metric is more expedient compared with GTF-IDF metric in case when the set of documents describe one subject domain. It can be explained by the fact that words, which are key for considered subject domain and occur in every documents of corpora, have a low IDF (as the result a low GTF-IDF). But in fact these words are informationally important and define the structure of the text.

4.2 Example 2

For comparison of the results of applying the visibility graph algorithm and the compactified horizontal visibility graph algorithm, the corpora of 2901 documents, which are related with an actual subject domain – “Complex network”, were considered from an open access repository of electronic preprints – arXiv (https://arxiv.org) for period of time 2000-2010. As the result of applying of visibility graph algorithms two different networks of words for considered subject domain, was obtained (Fig. 4, Fig. 5).

After deriving the associated graphs from the visibility algorithms, all the terms are sorted descending and weight values of CHVG and VG corresponding nodes according to a quantity of connections with other nodes are calculated. As the weight, for example, the authority (or hub) calculated by HITS algorithm [31] is used. Because of the graph is not directed, the choice of a form of the weight is not matter.

Comparing the results (Table 2), it may notice, that in the case of applying the compactified horizontal visibility graph algorithm (Fig. 5) there are many words, which have more links than in the case of applying the visibility graph algorithm (Fig. 4). A general quantity of links is 768 in the case of applying the compactified horizontal visibility graph algorithm, unlike in the case of applying the ordinary visibility graph algorithm, when a general quantity of links is 703. It should be noted, that obtained networks are very complex. That is why we plan to continue our research in this sphere.

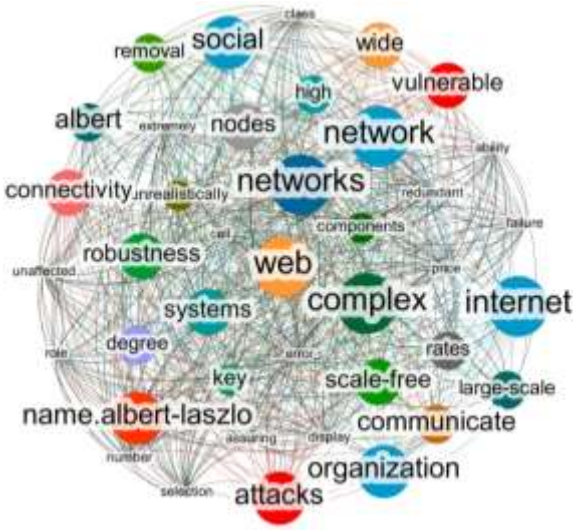


Fig. 4. The network of key words, obtained through the application of the visibility graph algorithm

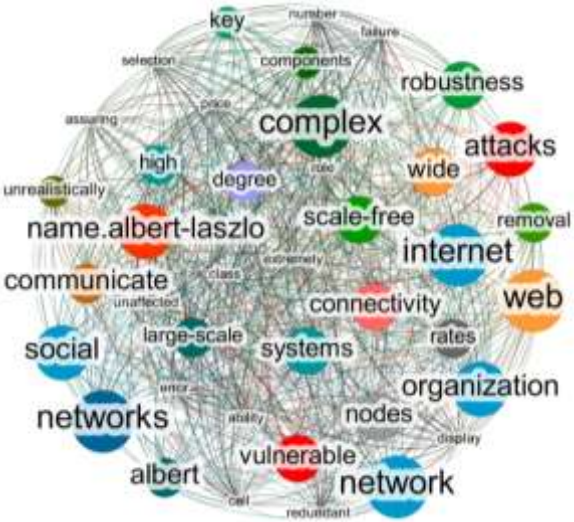


Fig. 5. The network of key words, obtained through the application of the compactified horizontal visibility graph algorithm

Table 2. TOP-40 largest-weight nodes of the network of words constructed from corpora of documents, which describe "Complex network" subject domain

Weight (VG)	Word	Weight (CHVG)	Word
0.17	networks	0.16	selection
0.17	web	0.16	networks
0.17	systems	0.16	removal
0.17	cell	0.16	systems
0.17	key	0.16	nodes
0.17	scale-free	0.16	network
0.17	display	0.16	cell
0.17	error	0.16	ability
0.17	complex	0.16	communicate
0.17	components	0.16	attacks
0.169	nodes	0.16	display
0.169	degree	0.16	rates
0.169	robustness	0.16	robustness
0.168	redundant	0.16	error
0.166	network	0.16	high
0.1645	rates	0.16	unaffected
0.1643	price	0.16	role
0.161	unrealistically	0.16	price
0.16	ability	0.16	connectivity
0.1582	extremely	0.16	extremely
0.158	number	0.158	web
0.1574	high	0.158	degree
0.1573	communicate	0.158	scale-free
0.156	assuring	0.158	wide
0.1547	social	0.158	unrealistically
0.1545	wide	0.157	number
0.1542	internet	0.157	complex
0.1509	unaffected	0.1573	vulnerable
0.1509	class	0.1573	class
0.1505	connectivity	0.1573	internet
0.1504	albert	0.1572	large-scale
0.1502	selection	0.1572	redundant
0.147	attacks	0.1539	social
0.146	removal	0.1539	failure
0.146	vulnerable	0.1534	key
0.145	role	0.1534	name.albert-laszlo
0.138	failure	0.1533	albert
0.135	name.albert-laszlo	0.1533	components
0.1015	large-scale	0.149	assuring
0.1012	organization	0.133	organization

5 Conclusion

The method to constructing networks from the texts, so-called language networks, was proposed. Key words and concepts from the set of documents which describe some subject domain were retrieved. Numeric values were assigned to each word or phrase using GTF metric, which was proposed in this article instead ordinary TF metric. After analyzing the results of research we made conclusion that applying only GTF metric is more expedient compared with GTF-IDF metric in case when the set of documents describe one subject domain. As the result a time series were constructed. A tool in time series analysis – the visibility graph algorithm was used for constructing the graph of subject domain. After analyzing the results of research the important structural elements of the text were found. It should be noted that these elements of the text also are informationally important and define the structure of the text. There was discovered, that in some case using the compactified horizontal visibility graph algorithm gives a network of words with more quantity of connections between concepts compared with using the visibility graph algorithm. Cause of complexity of obtained networks we plan to continue our research in this sphere.

The proposed method can be used for visualization some subject domain, and also for information support systems, enabling to reveal key components of the subject domain. Also the results of this article can be used for building UI of information retrieval systems, enabling to make a process of search a relevant information easier.

References

1. D.V. Lande, A.A. Snarskii, and I.V. Bezsudnov, Internetika: Navigation in complex networks: models and algorithms, Moscow, Russia: Librokom, Editorial URSS (in Russian) (2009).
2. D.V. Lande, Knowledge Search in INTERNET. Professional work. Dialectics, Moscow (in Russian) (2005).
3. C.C. Aggarwal, C.X. Zhai, Mining text data. Springer Science & Business Media (2012) 77-128.
4. G. Miner, J. Elder IV, and T. Hill, Practical text mining and statistical analysis for non-structured text data applications. Academic Press (2012).
5. V.Yu. Taranukha, Intelligent processing of texts, Kiev: electronic publication on the website of the faculty (in Ukrainian) (2014).
6. E.I. Bol'shakova, E. S. Klyshinsky, D.V. Lande, A.A. Noskov, O. V. Peskova, E.V. Yagunova, Automatic processing of texts in a natural

language and computational linguistics, Moscow: MIEM Publ (in Russian) (2011).
7. L. Lacasa, B. Luque, F. Ballesteros, J. Luque, J.C. Nuño, From time series to complex networks: the visibility graph, Proc. Natl. Acad. Sci. USA 105 (2008) 4972–4975.
8. A.M. Nunez, L. Lacasa, J. P. Gomez, Luque B. Visibility algorithms: A short review, Frontiers in Graph Theory. InTech, (2012) 119 – 152.
9. B. Luque, L. Lacasa, F. Ballesteros, J. Luque, Horizontal visibility graphs: Exact results for random time series. Physical Review E, 80(4) (2009) 046103.
10. G. Gutin, T. Mansour, S. Severini, A characterization of horizontal visibility graphs and combinatoris on words, Physica A, – 390 (2011) 2421-2428.
11. D.V. Lande, A.A. Snarskii, Compactified HVG for the Language Network. In: Proceedings of the International Conference on Intelligent Information Systems: The Conference is dedicated to the 50th anniversary of the Institute of Mathematics and Computer Science, 20-23 Aug. 2013, Chisinau, Moldova: Proceedings IIS, Institute of Mathematics and Computer Science (2013) 108–113.
12. D.V. Lande, A.A. Snarskii, E.V. Yagunova, E. Pronoza, The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text. In: Proceedings of the 12th Mexican International Conference on Artificial Intelligence (2013) 209–215
13. D.V. Lande, A.A. Snarskii, E.V. Yagunova, Application of the CHVG-algorithm for scientific texts. In: Proceedings of the Open Semantic Technologies for Intelligent Systems (OSTIS), February 20 – 22th, Minsk (2014) 199–204
14. D.V. Lande, A.A. Snarskii, D.Yu. Manko, The Model of Words Cumulative Influence in a Text. In: XVIII International Conference on Data Science and Intelligent Analysis of Information. Springer, Cham (2018) 249-256.
15. D.V. Lande, A.A. Snarskii, E.V. Yagunova, E. Pronoza, S. Volskaya, Hierarchies of Terms on the Euromaidan Events: Networks and Respondents Perception, 12th International Workshop on Natural Language Processing and Cognitive Science NLPCS 2015 127-139.
16. R. Ferrer-i-Cancho, R.V. Solé, The Small World of Human Language, Proceedings of the Royal Society of London B: Biological Sciences 268.1482 (2001) 2261-2265.
17. S.N. Dorogovtsev, J.F.F. Mendes, Language as an Evolving Word Web, Proceedings of the Royal Society of London B: Biological Sciences 268.1485 (2001) 2603-2606.

18. S.M.G. Caldeira, T.C. Petit Lobao, R.F.S. Andrade, A. Neme, J.G. Miranda, The network of concepts in written texts, Preprint physics/0508066 (2005).
19. R. Ferrer-i-Cancho, R.V. Solé, R. Kohler, Patterns in syntactic dependency networks, Physical Review E 69.5 (2004) 051915.
20. R. Ferrer-i-Cancho, The variation of Zipf's law in human language, The European Physical Journal B-Condensed Matter and Complex Systems, (2005) 249-257.
21. A.E. Motter, A.P.S. De Moura, Y.C. Lai, P. Dasgupta, Topology of the conceptual network of language, Physical Review E, 65(6) (2002) 065102.
22. M. Sigman, G.A. Cecchi, Global Properties of the Wordnet Lexicon, Proceedings of the National Academy of Sciences 99.3 (2002) 1742-1747.
23. I.V. Bezsudnov, A.A. Snarskii. From the time series to the complex networks: The parametric natural visibility graph, Physica A: Statistical Mechanics and its Applications 414 (2014) 53-60.
24. X. Li, M. Sun, C. Gao, D. Han, M. Wang, The parametric modified limited penetrable visibility graph for constructing complex networks from time series. Physica A: Statistical Mechanics and its Applications, 492 (2018) 1097-1106.
25. M. Wang, H. Xu, L. Tian, H. E. Stanley, Degree distributions and motif profiles of limited penetrable horizontal visibility graphs. Physica A: Statistical Mechanics and its Applications (2018).
26. M. Wang, A.L. Vilela, R. Du, L. Zhao, G. Dong, L. Tian, H. E. Stanley, Exact results of the limited penetrable horizontal visibility graph associated to random time series and its application. Scientific reports, 8(1) (2018) 5130.
27. M. Wang, A.L. Vilela, R. Du, L. Zhao, G. Dong, L. Tian, H. E. Stanley, Topological properties of the limited penetrable horizontal visibility graph family. Physical Review E, 97(5) (2018) 052117.
28. J.D. Ullman, Data Mining, Mining of massive datasets. Cambridge University Press. (2011) 1–17.
29. J. Beel, B. GIPP, S. Langer, C. Breitinger, Research-paper recommender systems: a literature survey, International Journal on Digital Libraries. 17(4), (2016) 305-338
30. K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, MCB University Press 60, (2004) 493-502.
31. J.M. Kleinberg, Authoritative sources in a hyperlink environment. Journal of the ACM JACM. 46 (5) (1999) 604–632.

МЕТОД РОЗПОДІЛУ ТАБЛИЦЬ РЕЛЯЦІЙНОЇ БАЗИ ДАНИХ РІВНОГО ОБ'ЄМУ ТА РІЗНИМИ ЙМОВІРНОСТЯМИ ЗВЕРТАННЯ ДО НИХ В ІНФОРМАЦІЙНО-ОБЧИСЛЮВАЛЬНІЙ МЕРЕЖІ АСУ

Ігор Субач, Олександр Чаузов
Інститут спеціального зв'язку та захисту інформації Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського", Україна
igor_subach@ukr.net

Проведено аналіз функціонування сучасних систем управління базами даних (СУБД), що функціонують в інформаційно-обчислювальних мережах (ІОМ) автоматизованих систем управління (АСУ). Зроблено висновок про залежність продуктивності функціонування ІОМ АСУ від методу розподілу інформаційного ресурсу, який застосовується в ній. Відзначено, що в основу методу доцільно покласти багаторівневу ієрархічну модель виділення інформаційного ресурсу. Відмічено, що велика кількість параметрів,які впливають на розподіл інформаційного ресурсу, а також розмаїтість показників якості при визначенні характеристик розподілу і труднощі їх зведення до єдиного критерію, досить ускладнюють методи розв'язання задачі розподілу інформаційного ресурсу. При цьому суть цієї задачі полягає у раціональному розміщенні реляційних таблиць БД по різних типах апаратно-програмних засобів (АПЗ). Це дає можливість скоротити часові витрати на обробку запитів, з огляду на характер оброблюваних даних. Сформульовано задачу розподілу мінімізації часу доступу до таблиць розподіленої реляційної бази даних (РРБД) однакового об'єму та різними ймовірностями звертання до них. Зроблено висновок про неможливість її розв'язання стандартними методами внаслідок нелінійності обмежень в її постановці. Запропоновано метод рішення сформульованої задачі, який базується на специфіці обмежень задачі та цільової функції. Суть запропонованого методу полягає у звуженні допустимих рішень на основі врахування нелінійності зв'язків в обмеженнях задачі та методики ранжування блоків, що запропонована авторами.

Ключові слова: *автоматизована система управління, інформаційно-обчислювальна мережа, розподілена реляційна база даних*.

1 Вступ

Аналіз функціонування систем управління базами даних (СУБД) інформаційно-обчислювальних мереж (ІОМ) автоматизованих

систем управління (АСУ) [1, 2] показує, що метод розподілу інформаційного ресурсу ІОМ АСУ для забезпечення функціонування складових частин системи у значній мірі визначає її продуктивність.

Під час організації та функціонування СУБД використовується багаторівнева система обробки та зберігання даних. Для цього при проектуванні системи або її модернізації створюється модель ієрархічного виділення інформаційного ресурсу, яка може розглядатися досить автономно та незалежно від взаємодії із зовнішніми абонентами. Така модель застосовується в системах, де для більшості функціонуючих транзакцій існує порівняно великий припустимий час реакції на зовнішні впливи й потрібні більші об'єми пам'яті для зберігання масивів даних і програм.

Кожний наступний рівень моделі ієрархічного виділення інформаційного ресурсу характеризується збільшенням часу доступу до інформації та зниженням вартості зберігання одиниці даних.

Зміна характеристик ресурсу кожного рівня безпосередньо впливає на продуктивність і ефективність роботи ІОМ АСУ в цілому. Для кожної ІОМ АСУ потрібно розв'язувати оптимізаційну задачу розподілу обмеженого інформаційного ресурсу з метою одержання мінімального значення узагальненого показника.

Велика кількість параметрів, що впливають на розподіл інформаційного ресурсу, а також розмаїтість показників якості при визначенні характеристик розподілу і труднощі їх поєднання до єдиного критерію, досить ускладнюють методи розв'язання задачі розподілу інформаційного ресурсу. Тому доцільно розглядати процес розподілу інформаційного ресурсу у вигляді ряду часткових моделей, які безпосередньо пов'язані з характеристиками збережених даних [5].

Виділимо ряд особливостей функціонування СУБД в ІОМ АСУ:

рівноймовірне звертання до деяких реляційних таблиць даних рівного об'єму внаслідок того, що час розв'язання задач ІОМ АСУ і періоди звертання до збереженої в БД інформації, є прогнозованими;

однократність завдання реляційних таблиць, причому відомо заздалегідь, що структура даних дозволяє мати ряд реляційних таблиць однакового об'єму для одного ієрархічного рівня пам'яті;

різний об'єм реляційних таблиць, тобто наявність реляційних таблиць різного об'єму, але однакової структури. У цьому випадку можлива декомпозиція різних за об'ємом реляційних таблиць на рівні.

Такий підхід застосовується при проектуванні СУБД, що відрізняються суворою періодичністю обробки інформації, яка є характерною для деяких ієрархічних рівнів підсистем комплексів задач, які не пов'язані з процесом управління.

Тому під моделлю розподіленої реляційної бази даних (РРБД) будемо розуміти модель, що характеризується реляційними таблицями рівного об'єму, причому ймовірності звертання до них є різними. Назвемо дану модель − модель РРБД з реляційними таблицями рівного об'єму та різними ймовірностями звертання до них.

При інтенсивному потоці запитів до СУБД фактична швидкодія виконання задач ІОМ АСУ у значній мірі визначається часом обробки кожного запиту. Операції по обробці запитів до різних типів апаратно-програмних засобів (АПЗ) можуть частково або повністю сполучатися за часом, тобто фактична швидкодія істотно залежить від обраного способу обробки реляційних таблиць бази даних (БД) різними типами АПЗ.

У сучасних РРБД при обробці великих інформаційних масивів швидкість обробки істотно залежить від розміщення реляційних таблиць, що описують однотипні об'єкти [1−4]. Відповідно до задач, які виконуються АСУ, при проектуванні складних запитів до РРБД великої інформаційної ємності, необхідно вкластися в задані часові границі.

При цьому суть задачі розподілу інформаційного ресурсу полягає у раціональному розміщенні реляційних таблиць БД по різних типах АПЗ. Це дає можливість скоротити часові витрати на обробку запитів, з огляду на характер оброблюваних даних.

2 Постановка задачі мінімізації часу доступу до таблиць розподіленої реляційної бази даних

Для запиту, який формується на основі інформації, що отримується з M таблиць РРБД обсягу W, з ймовірністю звертання до s-ї таблиці − p_s, $(s = \overline{1, M})$, $\left(\sum_{s=1}^{M} p_s = 1 \right)$, у випадку моделі розподілу реляційних таблиць БД рівного об'єму з різними ймовірностями звертання до них [4, 5], сумарний час доступу до реляційних таблиць складе:

$$T = \sum_{i=1}^{M} \sum_{k=1}^{K_C} p_i \cdot x_{ik} \cdot \tau_k \,, \qquad (1)$$

де x_{ik} – булева змінна розподілу таблиць РРБД:

$$x_{ik} = \begin{cases} 1, \text{якщо i-а таблиця розміщується в k-му блоці;} \\ 0, \text{якщо i-а таблиця не використовує k-й блок.} \end{cases}$$

При обмеженнях:

кожна реляційна таблиця обробляється тільки в одному блоці, що не впливає на загальність постановки задачі, тому що $V_j > W$ $\forall j \in \overline{1,N}$:

$$\sum_{k=1}^{K_C} x_{ik} = 1, \quad i = \overline{1,M}; \qquad (2)$$

де $K_C = \sum_{j=1}^{N} n_j$ – загальна кількість доступних блоків;

кількість задіяних блоків типу j не повинна перевищувати максимально можливу кількість блоків даного типу, що необхідно з визначення змінних y_j:

$$y_j \le n_j, \quad j = \overline{1,N}; \qquad (3)$$

наведені сумарні витрати на необхідне число блоків АПЗ не повинні перевищувати максимально припустимого розміру витрат F, що задається нерівністю:

$$\sum_{j=1}^{N} f_j y_j \le F; \qquad (4)$$

змінні задачі повинні належати заданій області та бути цілочисельними:

$$x_{ik} \in \{0,1\}, \; y_j \in \{0,1,...,n_j\}. \qquad (5)$$

Вимога про достатність числа блоків типу j для обробки реляційних таблиць БД записується нерівністю:

$$\sum_{i=1}^{M} \sum_{k=\varphi(j-1)+1}^{\varphi(j-1)+y_j} x_{ik} \le l_j \cdot y_j \qquad \forall\, j \in \overline{1,\mathrm{N}}, \tag{6}$$

де $l_j = [V_j / W]$ − число реляційних таблиць, які можуть оброблятися в одному блоці j-го типу;

$\varphi(j)$ – функція зміщення номеру на множині $\{0, ..., N\text{-}1\}$ типів обчислювальних вузлів:

$$\begin{cases} \varphi(0) = 0; \\ \varphi(j-1) = \sum\limits_{k'=1}^{j-1} n_{k'}, \quad \mathrm{j} = \overline{2,\mathrm{N}}. \end{cases} \tag{7}$$

Функція зміщення номера дозволяє для обчислювального вузла типу j визначити порядковий номер першого з блоків, які представлені для запиту $k_j^{(1)} = \varphi(j\text{-}1)+1$.

З урахуванням цього, число реляційних таблиць, що розміщені у блоках типу j, дорівнює:

$$\left[\left(\sum_{i=1}^{M} \sum_{\mathrm{k}=\varphi(\mathrm{j}-1)+1}^{\varphi(\mathrm{j}-1)+\mathrm{y_j}} \mathrm{x_{ik}} \right) / l_j \right]. \tag{8}$$

Вимога щодо мінімізації часу доступу до розподілених таблиць БД при обробці та розміщенні РРБД приводить до задачі цілочисельного нелінійного програмування з цільовою функцією:

$$T = \sum_{i=1}^{M} \sum_{k=1}^{K_C} p_i \cdot x_{ik} \cdot \tau_k \to \min \tag{9}$$

та обмеженнями (2) – (7).

Наявність змінної y_j у межах суми обмеження (6) не дозволяє розв’язати задачу цілочисельного програмування (9) стандартними методами, оскільки дане обмеження не є лінійним.

Тому для рішення даної задачі необхідно використовувати методи, що базуються на специфіці обмежень та цільовій функції [7].

3 Метод розподілу таблиць розподіленої реляційної бази даних в ІОМ АСУ

Аналіз публікацій [8−11] показує, що незважаючи на те, що задача цілочисельного програмування має значну складність щодо її

вирішення, для неї розроблено достатньо велику кількість алгоритмів. Деякі з цих алгоритмів ефективні для окремих класів задач цілочисельного програмування, однак, у загальному випадку можна стверджувати, що не існує загального алгоритму, який би знаходив оптимальне рішення за достатній час для задач великих розмірностей.

Тому для рішення оптимізаційної задачі (9) при обмеженнях (2−7) пропонується використовувати метод звуження області допустимих рішень, що базується на урахуванні нелінійного зв'язку змінних в обмеженнях (2−7) та методиці ранжування блоків, що була запропонована авторами.

Перший етап пропонованого методу полягає в розбивці області допустимих рішень на дві підмножини X та Y, де Y − цілочисельна множина векторів розподілу блоків та X − множина планів розподілу реляційних таблиць.

Кожний елемент множини допустимих рішень задачі (2−9) $Z = \{z = (x_{11}, ..., x_{ik}, ..., X_{MK_C}, y_1, ..., y_j, ..., y_N)\}$ можна представити як конкатенацію плану розподілу реляційних таблиць по блоках вузлів ІОМ з множини $X = \{x = (x_{11}, ..., X_{MK_C})\} \subset \mathrm{B}^{M \times K_C}$ та вектора розподілу задіяних блоків з множини: $Y = \{y = (y_1, ..., y_J)\} \subset R^N : z = x \,!!\, y$.

Кожен фіксований вектор розподілу y_{fix} однозначно визначає план розподілу, виходячи з обраного способу впорядкування блоків:

$$\begin{aligned} &x_{ik} = 1 \text{ для } i = (k-1)\cdot \mathrm{l_j} + 1, ..., k\cdot l_j,\ i \le M; \\ &k = \varphi(j-1) + 1, ..., \varphi(j-1) + y_j,\ j = \overline{1,N}; \\ &x_{ik} = 0 \text{ для } i = (k-1)\cdot \mathrm{l_j} + 1, ..., k\cdot l_j,\ i \le M; \\ &k \ne \varphi(j-1) + 1, ..., \varphi(j-1) + y_j,\ j = \overline{1,N}; \end{aligned} \qquad (10)$$

На другому етапі з множини векторів розподілу блоків вузлів видаляються всі елементи, що не задовольняють умові (2). Потім, використовуючи різні варіанти ранжування по типах вузлів, корегуються верхня та нижня межа розбивки множини Y по рівнях задіяних блоків таким чином, щоб в отриманій підмножині Y^* містився вектор оптимального розподілу блоків $y^{(\text{опт})}$ та $card\ Y^* << card\ Y$.

Уведемо на множині Y бієктивне відображення на підмножині цілих невід'ємних N-розрядних p-х чисел**:**

$$\psi: Y \to Q_p, \qquad (11)$$

так, що $\psi(y) = r_p = \left(\sum_{j=1}^{N} y_j \cdot p^{j-1} \right)_p$, де $p = 1 + \max\limits_{j \in \overline{1,N}} n_j$.

З обмеження (2) та бієктивності введеного обмеження можна зробити висновок, що потужність множини Q_p обмежена зверху нерівністю:

$$card Q_p \le \prod_{j=1}^{N} (n_j + 1). \qquad (12)$$

Очевидно, що існує єдине розбиття Ω множини Q_p на K_C+1 підмножин, які не перетинаються та в кожному з яких є фіксована сума p-ічних цифр:

$$Q_p = \bigcup_{s=0}^{K_c} Q_p^{(s)}, \qquad (13)$$

де $Q_p^{(s)} = \left\{ r_p \mid \sum_{j=1}^{N} y_j = s, s = \overline{0, K_c} \right\}$, (14)

причому нижньою межою Ω є $Q_p^{(0)}$, а верхньою – $Q_p^{(K_C)}$.

Відмітимо, що серед елементів розбиття Ω знайдуться такі $Q_p^{(\zeta_{\min})}$ та $Q_p^{(\zeta_{\max})}$, що:

$$\zeta_{\min}, \zeta_{\max} \in \overline{1, K_c} : \zeta_{\min} \le \sum_{j=1}^{N} y_j \le \zeta_{\max}. \qquad (15)$$

У результаті до множини Q_p^* потрапляють тільки ті p-ічні числа из Q_p, в яких сума цифр знаходиться у межах від ζ_{min} до ζ_{max}, у тому числі й $r_p^{(onm)}$.

На третьому етапі з підмножини Y^* гіперплощинами відсікаються елементи, що не вдовольняють обмеженню (3) задачі та визначається підмножина $Y^{**} \subset Y^*$, що містить $y^{(опт)}$.

Для подальшого зменшення потужності множини Q_p^* врахуємо специфіку обмеження (2) задачі, що розглядається.

Покажемо, що $\exists \tilde{j} \le N, \exists \underset{\sim}{\zeta}_j \in \overline{0, K_c} : \underset{\sim}{\zeta}_j \le \sum_{j=1}^{\tilde{j}} y_j$.

Послідовно знаходячи верхню межу суми елементів Q_p^*, отримаємо підмножину Q_p^{**}, що містить число $r_p^{(onm)}$. Застосовуючи операцію зворотнього відображення ψ^{-1} до елементів множини Q_p^{**} можна отримати підмножину векторів розподілу блоків Y^{**}:

$$\left(\left(Y^{**} \subset Y^*, y^{onm} \in Y^{**}\right), \text{оскільки } \psi^{-1}\left(r_p^{(onm)}\right) = y^{onm}\right).$$

Оптимальний план $y^{(\text{опт})} \subset Y^{**}$ розподілу реляційних таблиць по блоках вузлів ІОМ визначається на четвертому етапі за допомого введеної функції зміщення номера (7).

Використовуючи (10) визначаємо план оптимального розподілу інформаційного ресурсу по блокам вузлів – $x^{(\text{опт})}$. Тоді рішенням задачі є вектор $z^{(onm)} = \left(x^{(onm)}\right) !! \left(y^{(onm)}\right)$.

Запропонований метод, що заснований на врахуванні специфіки змінних та обмежень оптимізаційної задачі можливо використовувати для реалізації більш складних моделей розподілу інформаційного ресурсу.

Джерела

1. Кульба В.В. Теоретические основы проектирования оптимальных структур распределенных баз данных / [Кульба В.В., Ковалевский С.С., Косяченко С.А., Сиротюк В.О.] // Серия «Информатизация России на пороге XXI века». – М.: СИНТЕГ, 1999. – 660 с.
2. Ульман Дж. Основы систем баз данных / Джон Ульман : [пер. с англ. М.Р.Когаловского и В.В.Когутовского]. – М.: Финансы и статистика, 1983. – 334с..
3. Жожикашвили В. А., Вишневский В. М. Сети массового обслуживания. Теория и применение к сетям ЭВМ. - М.: Радио и связь, 1988. – 189 с.
4. Логинов И.В. Оптимизация модели распределенной гетерогенной вычислительной системы, используемой для планирования обработки запросов [Текст] / И.В. Логинов, Е.В. Лебеденко // Информатика и системы управления. – 2009. – №3(21). – С. 118-124.

5. Субач І.Ю. Моделі розподілу інформаційного ресурусу в АСУ спеціального призначення // І.Ю. Субач, О.М. Чаузов, Н.Г. Кучук // Information Technology and Security. – 2016. – Vol 4., Iss. 1. – P. 74–83.
6. Чаузов О.М. Математична модель розподілу інформаційного ресурсу між транзакціями до сховищ даних / О.М. Чаузов // Системи управління навігації та зв'язку, Полтава: 2015. – Випуск 4(36). – С. 100–102.
7. Субач І.Ю. Метод рішення задачі розподілу інформаційного ресурсу в АСУ спеціального призначення при варіативному розмірі інформаційних блоків // І.Ю. Субач, О.М. Чаузов, Н.Г. Кучук // Information Technology and Security. – 2016. – Vol 4., Iss. 2. – P. 269–276.
8. Хемди А.Таха. Введение в исследование операций [7–е изд.; пер. с англ] / Хемди А.Таха. – М.:Издательский дом «Вильямс», 2005. – 912 с.
9. Вентцель Е. С. Исследование операций / Е.С.Вентцель. – М.: Советское радио, 1972. – 552 с.
10. Тжаскалик, Т. Введение в исследование операций с применением компьютера: Пер. с польск. И. Д. Рудинского. / Т. Тжаскалик - Москва : Горячая линия - Телеком, 2009. - 440 с.
11. Гвишиани Д.М. Многокритериальные задачи принятия решений [Текст] / Д.М. Гвишиани; под ред. Д.М. Гвишиани и С.В. Емельянова. – М.: Машиностроение, 1978. – 192 с : ил.

МЕТРИЧНИЙ ПІДХІД ДО ОЦІНКИ РИЗИКУ КІБЕРАТАК НА ГЛОБАЛЬНУ МАРШРУТИЗАЦІЮ

Віталій Зубок
Інститут проблем моделювання в енергетиці
ім. Г.Є.Пухова НАН України
Вул. Генерала Наумова, 15, Київ, 03164, Україна
vitaly.zubok@gmail.com

Одією з масштабних проблем кібербезпеки є запобігання перехопленню маршрутів в системі глобальної маршрутизації мережі Інтернет. На основі сучасної світової практики поводження з ризиками пропонуються теоретичні засади ідентифікації та оцінки ризиків перехоплення маршруту через дослідження топології зв'язків між автономними системами мережі Інтернет для подальшого формулювання задачі обробки ризиків як задачі про топологію.
Ключові слова: *глобальна маршрутизація, перехоплення маршрутів, оцінка ризиків, кібербезпека.*

1 Актуальність

В роботі [1] пригорнуто увагу до масштабу загроз, пов'язаних з атаками на Інтернет-маршрутизацію, та необхідності всебічного аналізу даної проблемної області з метою пошуку методів зменшення впливу таких атак, які матимуть важливе значення для кіберзахисту як на корпоративному рівні, так і на рівні критичної інфраструктури держав.

Один із напрямків визначено як необхідність запобігання перехопленню маршрутів до власних префіксів. Напрямок дослідження сформульовано як задачу пошуку найбільш ефективної топологічної організації зв'язків на рівні глобальної маршрутизації в мережі Інтернет, що забезпечить мінімізацію втрат від перехоплення маршруту в межах певної цільової групи вузлів. Метою даної роботи є визначення зв'язку між топологією та ризиком перехоплення маршруту.

2 Ризик перехоплення маршруту в термінах та визначеннях міжнародних стандартів

В сучасній світовій практиці поводження з ризиками існує основа єдиного методичного підходу до сприйняття документів, які

регламентують різні аспекти діяльності. Такою основою є настанови ISO Guide 73:2009 "Risk Management – Vocabulary", які тлумачать зміст відповідних термінів [2]. Головним є поняття ризику, яке надано як вплив невизначеності на досягення цілі або мети. Проте, оскільки таке поняття ризику неможливе у знеособленому сенсі, важливо насамперед визначити, хто є зацікавленою стороною (stakeholder) в оцінюванні ризику. В даній роботі такою стороною є суб'єкт глобальної маршрутизації, оскільки в наслідок можливого перехоплення маршруту саме він отримає збитки.

Ризик може матеріалізуватись як настання потенційно можливих подій та (або) наслідків цих подій. Значення ризику можна виразити як поєднання подій (і їхніх наслідків) із вірогідністю їх настання. Такою подією вважатимемо свідомі чи несвідомі дії третіх сторін, які призвели до такого наслідку, як несанкціонована поява в мережі альтернативних, більш пріоритетних маршрутів.

Ризик має аналізуватись у контексті оточення, яке поділяється на зовнішнє та внутрішнє. До внутрішнього оточення спробуємо віднести внутрішню політику маршрутизації, а о зовнішнього оточення – весь процес глобальної маршрутизації в цілому, який полягає у відносинах зацікавленої сторони з усіма іншими суб'єктами глобальної маршрутизації. Ці відносини матеріалізуються, зокрема, в обміні маршрутами по протоколу BGP-4 та в інтерпретації (сприйнятті) глобальної таблиці маршрутизації.

Оцінювання ризику потребує, серед іншого, його ідентифікації. Оскільки ризик обумовлений особливостями зовнішнього і внутрішнього середовища, розглядаються всі можливі джерела ризику, а також наявна інформація про сприйняття ризику (усвідомлення ризику) причетними сторонами, як внутрішніми по відношенню до компанії, так і зовнішніми. Особливі вимоги висуваються до якості інформації (максимально можливий рівень повноти, точності і тимчасової відповідності при наявних ресурсах на її отримання) та її джерел. Результат ідентифікації повинен бути структурованим та охоплювати чотири елементи – джерела виникнення; події, що виникнуть; причини цих подій; наслідки подій. Для ідентицікації ризику перехоплення маршруту зробимо опис цього ризику на основі дослідження відомих тактик та стратегій таких атак перехоплення маршрутів [1] та узагальнимо цю інформацію. Отже:

– джерелами виникнення ризику обов'язково є інші суб'єкти глобальної маршрутизації;

– події, виникнення яких спричинює ризик, це несанкціоновані зміни в глобальній таблиці маршрутизації чи її інтерпретації на інших суб'єктах глобальної маршрутизації;
– наслідками цих подій є несанкціонована зміна напрямку проходження мережевого трафіку.

3 Загальний підхід до оцінки ризику перехоплення маршруту

Як відомо з принципів організації глобальної маршрутизації та протоколу BGP-4, основним транзитивним параметром, що характеризує привабливість маршруту, є довжина шляху (AS_PATH) [3]. Довжина шляху – це фактор, який дозволяє маршрутам до однакових префіксів конкурувати. Інтернет на цьому рівні являє собою незважений граф, вершинами якого є автономні системи. В загальному випадку граф є циклічним та обов'язково зв'язним. Математично цей граф можна представити або квадратною матрицею суміжності, або квадратною матрицею відстаней розмірності N, де N – кількість вузлів [4].

Якщо існує підмножина вузлів, об'єднана якоюсь сутністю, топологія цієї підмножини може розглядатись окремо. Назвемо цю підмножину цільовою групою вузлів. Такою групою можуть бути вузли – учасники будь-якої мережі обміну трафіком чи вузли-клієнти одного провайдера доступу до Інтернет.

Якщо зловмисник вдало провів перехоплення маршруту чи захоплення префіксу, це означає, що для певної цільової групи вузлів маршрут до префікса жертви через вузол зловмисника став коротшим, ніж інші, природні маршрути, а отже - буде перехоплено трафік до цього префіксу від згаданої групи вузлів.

Як вже згадувалось, у сучасній практиці для формалізації ризику широко використовують моделі, які пов'язують між собою ймовірність виникнення негативних подій і можливих збитків у результаті цих подій [5]. Визначимо ризик перехоплення трафіку R до певного префіксу як добуток ймовірності P такого перехоплення та збитку C, пов'язаних з цим перехопленням. Збиток є в свою чергу сумою збитків від перехоплення трафіку від кожного з вузлів в цільовій групі, тому:

$$R = P\sum_i C_i \quad . \qquad (1)$$

Якщо розподіл збитків між вузлами заздалегідь невідомий, виправданим буде вважати його однаковим для кожного вузла. Тоді збиток є пропорційним до кількості вузлів в цільовій групі. Тоді можливо оцінювати ризик як величину, пропорційну кількості вузлів *N*, що потрапили під вплив перехоплення:

$$R = NC. \qquad (2)$$

4 Метричний підхід до визначення ризику

Проаналізуємо, від чого залежить ймовірність перехоплення трафіку P. Перехоплення означає, що маршрут до префікса жертви через вузол зловмисника став коротшим, ніж істинний маршрут. Існує поняття метричної розмірності графа (metric dimension) - такої мінімальної кількості вузлів графа, що положення інших вузлів може бути точно описано відстанями до перших. Відстань між вузлами як довжина найкоротшого маршруту для мережі Інтернет [6] - це функція:

$$d(v,u) = \min_i \left(d(v,i) + d(i,u)\right). \qquad (3)$$

З практичної точки зору це означає, що в разі перехоплення маршруту відстань (3) через фіктивний маршут стане меншою, ніж через справжній маршрут. Маніпулювати довжиною шляху тим простіше, чим цей шлях довший (в довшому шляху посередині існує більше вузлів, через які можна анонсувати фіктивний маршрут). Отже, ймовірність перехоплення $P(v,u)$ між вузлами *v,u* збільшується для далеких вузлів та зменшується для близьких:

$$P(v,u) \sim d(v,u). \qquad (4)$$

Отже, ризик пов'язаний з кількістю вузлів, що можуть потрапити під вплив перехоплення і з відстанню до кожного з цих вузлів.

В роботі [6] представлено дослідження Інтернету з точки зору теорії складних мереж та було показано зв'язок між середнім шляхом мережі, її ефективністю та вразливістю. Для кожного конкретного вузла *v* за відомими відстанями $d(v,i)$ можна визначити суму відстаней:

$$D_v = \sum_{i=1}^{|V|} d(v,i) \qquad (5)$$

Застосовуючи (4) до множини вузлів V, з урахуванням (5) можна отримати залежність ризику перехоплення маршрутів до вузла v від його положення відносно інших вузлів:

$$R_v \sim \sum_{i=1}^{|V|} d(v,i)\,. \qquad (6)$$

5 Висновок

З використанням метричної функції для мережі Інтернет, яка представлена у вигляді графа, встановлено зв'язок між положенням вузла в мережі та ризиком перехоплення маршрутів до нього. Це дає в подальшому можливість сформулювати задачу керування ризиками для певного вузла від перехоплення маршрутів як задачу пошуку для нього найбільш ефективної топології зв'язків.

References

1. Зубок, В: Визначення напрямків протидії кібератакам на глобальну маршрутизацію в мережі Інтернет. Електронне моделювання 55(40), 10-15 (2016).
2. Risk Management – Vocabulary (ISO Guide 73:2009, IDT) : ДСТУ ISO Guide 73:2013. – [Чинний від 2014–07–01] . – Київ : Мінекономрозвитку України, 2014. – 13 с. - (Національні стандарти України).
3. Rekhter, Y. and Li, T. and Hares, S.: A Border Gateway Protocol 4 (BGP-4). https://tools.ietf.org/html/rfc4271. Дата доступу: 29 червня, 2018 р.
4. Зубок, В.: Практические аспекты моделирования изменений в топологии глобальных компьютерных сетей. Реєстрація, зберігання і обробка даних 2(14), 67-78 (2012).
5. Joint Task Force Transformation Initiative. Guide for Conducting Risk Assessments. NIST Special Publication 800-30 (2012).
6. Мохор, В. та Зубок, В. Формування міжвузлових зв'язків в Інтернет з використанням методів теорії складних мереж. К.:«Прометей», 2017. – 175с.

METHOD OF FORMING THE RING CODES

Otrokh S.I.[1], Kuzminykh V.O.[2], Hryshchenko O.O.[1]
[1]*State University of Telecommunications, Kyiv, Ukraine,*
[2]*National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine, vakuz0202@gmail.com*

The article considers the method of forming a type of cyclic code - the ring code. The purpose of creating the ring code was the compression of information and its protection against unauthorized access. The article presents the structure of the forming matrix and the mathematical model for the formation of the ring code. To identify the ring code the shift indexes vector is proposed. An algorithm for constructing a shift indexes vector is given on the example of a specific ring code. The properties of shift indexes vector, created by summing the number of units obtained as a result of one of the binary transformations of the XOR, OR, AND elements of the initial sequence (first line) of the ring code and successively on each subsequent line, are investigated. The formulas of calculation of the sum of the decimal values of the elements of the shift indexes vector are given. This method can be used to build an effective channel for the transmission of the future network. The result of the study determined that the compression of information on the use of the ring code is 2.7 times compared with the amount of information transmitted.

1 Introduction

Ring codes are built on the principle of block cyclic codes, the rows of forming matrices which are interconnected as a condition of cyclicity. The offset of the elements of the code sequence of the cyclic code is done from right to left, and, the leftmost symbol is always transferred to the right at the end of the code sequence. Each line of the forming matrix of a cyclic code has the same number of elements and the same structure of combinations of units and zeros, but the number of rows and columns in the forming matrix can be arbitrary. A ring code is a kind of cyclic code and, unlike the latter, is always a square matrix of size $N \times N$, in which the number of rows corresponds to the number of columns. Each line of the forming matrix of the ring code contains m units and, accordingly, N -

m zeros. In this case, the rows of the matrix form a ring of a complete cycle of shift of elements of the code sequence.

2. The algorithm of the ring codes formation

The algorithm for the formation of a ring code is shown in Fig. 1.

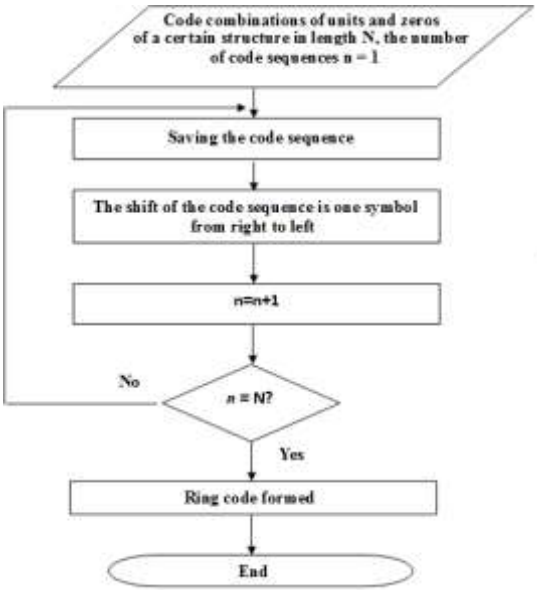


Fig. 1. The algorithm of the ring code formation

In general, the code sequence of the ring code can be determined as:

$$C(x) = k \cdot x^{N-1} \ldots + \cdots k \cdot x^2 + k \cdot x^1 + k \cdot x^0, \qquad (1)$$

where k – coefficient that acquires the value of 1 or 0, and x –2 (the basis of the binary number system), and 0,1, ... N-1 – is the bit number of the binary system of the number.

Then the forming matrix of a circular code of size $N \times N$ in the general form is formed so that the number of rows equals the number of columns, that is, the matrix has a square shape. In this case, the matrix can be recorded that (2):

$$C(N,N) = \begin{bmatrix} k \cdot x_1^{N-1} \ldots + \cdots k \cdot x_1^2 + k \cdot x_1^1 + k \cdot x_1^0 \\ k \cdot x_2^{N-1} \ldots + \cdots k \cdot x_2^2 + k \cdot x_2^1 + k \cdot x_2^0 \\ \vdots \\ k \cdot x_N^{N-1} \ldots + \cdots k \cdot x_N^2 + k \cdot x_N^1 + k \cdot x_N^0 \end{bmatrix} \quad (2)$$

3 The mathematical model of the ring code formation

The process of forming a ring code possible to describe mathematically. The binary code sequences of the ring code can be represented in the decimal system in the form of their decimal values, the mathematical model for the formation of which takes on this form:

$$C_i(N,m) = S_1 \cup S_2, \quad (3)$$

where $C_i(N,m)$ – total set of decimal values of code sequences, S_1 is a set of a first decimal values of the code sequences of the ring code, S_2 – a set of a second code decimal values of the ring code sequences, N – the number of code sequences of the ring code equal to the number of elements of the code sequence, m – the number of units in code sequence.

At that

$$S_1 = \{s_{11}, s_{12}, \ldots, s_{1n)}\},\ S_2 = \{s_{21}, s_{22}, \ldots, s_{2l}\} \quad (4)$$

where: S_{11} – the decimal value of the first code sequence of the set 1, S_{1n} – the decimal value of the last code sequence of the set 1, S_{21} – the decimal value of the first code sequence of the set 2, S_{2l} – the decimal value of the last code sequence of the set 2.

Common expressions for computing the decimal values of the elements set S_1 are as follows:

$$s_{11} = \sum_{i=0}^{N-1} k * 2^i \quad (5)$$

where k – coefficient that acquires the value of 1 or 0, N is the number of elements of the code sequence, and s_{11} - the least decimal value of the code sequence.

It should be noted that the decimal value of each subsequent code sequence is twice the decimal value of the previous code sequence of the set S_1. Therefore:

$$s_{12} = 2 \cdot s_{11};\ s_{1n} = 2 \cdot s_{1(n-1)}, \quad \backslash \quad (6)$$

where n – the number of code sequences in a set S_1.

The general expressions for computing the decimal values of the set S_2 are as follows:

$$s_{21} = s_{1n} - S_{\text{p}};\ s_{22} = 2 \cdot s_{21};\ s_{2l} = 2 \cdot s_{1(l-1)}. \quad (7)$$

Where S_p – the difference, the calculation formula of which depends on the structure of the combinations of one and zero symbols of the code sequences of the ring code; l - the quantity of code sequences in a set 2.

The ring code is characterized by a vector of shift indices (VSI), which is formed by summing the number of units obtained as a result of one of the binary transformations *XOR*, *OR*, *AND* (with N_{ot} or without it) of the elements of the initial sequence (first line) of the ring code and successively each of the next line.

For example, the shift indexes vector of the ring code of 7 × 7, each line contains 4 units and 3 zeros, and the initial vector consists of a code sequence [0101011], is {624426}, as shown in table.1. Moreover, the quantity of symbols in the shift indexes vector per unit is less than the number of lines of the ring code. It should also be noted that the shift indexes vector is a group integral index of the whole ring code, rather than a single line of it.

Thus, in the communication channel, we can transfer 49 binary information symbols, unless we use the ring code. If we use ring code in the communication channel, we transmit instead of 49 symbols only 18. Due to the use of the ring code, avoid redundancy and the gain is 49/18 = 2.7 times.

Table 1. The process of creating the ring code

Formation matrix	The matrix of shift indexes vector in the binary system as a result of the logical operation XOR	Shift indexes vector in decimal (number of units)	The matrix of shift indexes vector in in a binary system	Shift indexes vector in a binary system
0101011	1111101	6	110	
1010110	0000110	2	010	
0101101	1110001	4	100	
1011010	1110001	4	100	110010100100010110
0110101	1000001	2	010	
1101010	1111110	6	110	
1010101				

The mathematical expressions of the formation of the vector of shift indices (VSI) by summing up the number of units obtained as a result of the implementation of the binary transformations *XOR*, *OR* and *AND*, respectively, become as follows:

$$VSI_{XOR} = \sum_{i=1}^{N}(x_{1i}\ XOR\ x_{2i}) \cup \sum_{i=1}^{N}(x_{1i}\ XOR\ x_{3i}) \ldots \cup \ldots \sum_{i=1}^{N}(x_{1i}\ XOR\ x_{Ni}), \quad (8)$$

$$VSI_{OR} = \sum_{i=1}^{N}(x_{1i}\ OR\ x_{2i}) \cup \sum_{i=1}^{N}(x_{1i}\ OR\ x_{3i}) \ldots \cup \ldots \sum_{i=1}^{N}(x_{1i}\ OR\ x_{Ni}), \quad (9)$$

$$VSI_{AND} = \sum_{i=1}^{N}(x_{1i}\ AND\ x_{2i}) \cup \sum_{i=1}^{N}(x_{1i}\ AND\ x_{3i}) \ldots \cup \ldots \sum_{i=1}^{N}(x_{1i}\ AND\ x_{Ni}), \quad (10)$$

where $x_{1i}, x_{2i}, x_{3i}, \ldots, x_{Ni}$ i-th element 1-st, 2-nd, 3-th, *N*-th lines of the ring code.

As a result of research of the structure of VSI formed through the binary transformations XOR, OR, AND, the following patterns were found:

— the elements of any VSI are placed symmetrically with respect to its center;

— the sum of the decimal value of the element formed by the binary XOR transformation, and the decimal value of the element formed by the binary transformation AND, is equal to the decimal value of the element, formed by binary OR;

— vectors of the indices of the shift of the ring code can be obtained both in rows and in the columns of the matrix of the ring code;

— the structure of the XOR vector of shift indices remains unchanged if the value of the symbols of the code sequence of the ring code changes to the opposite. The AND and OR vectors of the displacement indices do not have this property.

The sum of the decimal values of the VSI elements formed by the OR-transformation consists of the sum of the decimal values of the VSI elements formed by the XOR transformation and from the sum of the decimal values of the VSI elements generated by the AND transformation. At the same time, the analysis of the structure of the vector of shift indices and their total values allows to note that, regardless of the number of elements of *N* and the number of single symbols m in the code sequence, there is a functional dependence between the sum of the decimal values of the elements of the shift indexes vector and the number of zero and single symbols. It represented by the following formulas:

1) for VSI, created by the XOR-transformation:

$$S_{VSI(XOR)} = (N - m) \cdot 2m, \quad (11)$$

where $S_{VSI(XOR)}$ – the sum of decimal values elements of the VSI generated by the binary XOR-transformation.

In order to determine the number of one and zero symbols in the code sequence, you can apply the formula for calculating the discriminant and the roots of the quadratic equation:

$$x_{1,2} = \frac{N \pm \sqrt{N^2 - 4\frac{S_{VSI}(XOR)}{2}}}{2}, \qquad (12)$$

where $x_{1,2}$ - the quantity of one and zero symbols m and (N - m), N - the length of the code, $S_{VSI(XOR)}$ – the sum of the decimal values of the elements VSI generated by the binary XOR-transformation.

2) for VSI, created by binary AND-transformation:

$$S_{VSI(AND)} = (m-1) \cdot m, \qquad (13)$$

where $S_{VSI(AND)}$ – the sum of the decimal values of the VSI elements generated by the binary AND-transformation.

In order to determine the number of one and zero symbols in the code sequence, you can apply the following simple formulas

3) for VSI, created by the binary OR-transformation:

$$S_{VSI(OR)} = (N-m) \cdot 2m + (m-1) \cdot m = N \cdot 2m - 2m^2 + m^2 - m = \\ = N \cdot 2m - m^2 - m = m \cdot (2N - m - 1), \qquad (14)$$

where $S_{VSI(OR)}$ – the sum of the decimal value VSI elements generated by the binary OR-transformation.

In Table 2 and on the Fig. 2 shows the dynamics of change in the sum of the decimal value VSI elements generated by the binary transformations XOR, OR, AND, depending on the number of single elements m of the code sequence of ring code.

Table 2. The dynamics of change in the sum of the decimal values of the shift indexes vector elements, depending on the number of elements of N and the number of single symbols of m code sequences

Number of units m	The sum of the decimal values of the elements VSI		
	XOR- VSI	AND- VSI	OR- VSI
The number of items N=7			
1	12	0	12
2	20	2	22
3	24	6	30
4	24	12	36
5	20	20	40
6	12	30	42

4 Conclusion

The method of the ring code generation was developed and the mathematical models of forming the ring code families for the construction of an effective channel for the transmission. The mathematical models of forming the ring code families were developed with using the values of code sequences in the decimal system.

The vector of shift indices, created by summing the number of units obtained as a result of one of the binary transformations of the XOR, OR, AND elements, was developed. The shift indexes vector is analog of the ring code, which can be transmitted via a communication channel instead of code.

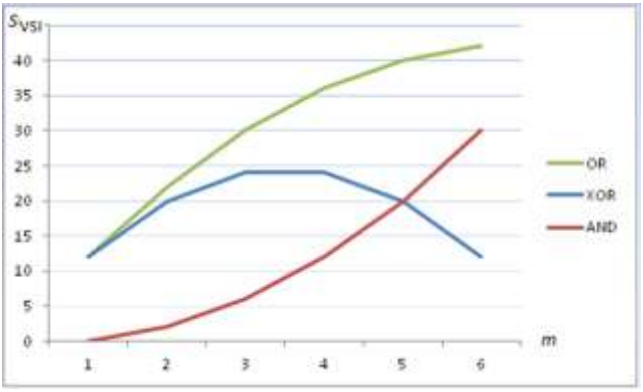


Fig. 2. Dynamics of the change of the sum of the decimal values of the elements of the VSI, created by the binary transformations XOR, OR, AND, from the number of single elements m for a 7x7 ring code

Formulas for determining the sum of the decimal values of the e shift indexes vector elements, obtained by the AND, OR and XOR transformation, were derived. It was determined that there is a functional dependence between the sum of the decimal values of the elements of the shift indexes vector and the number of zero and single symbols in the code sequence of ring code.

The dynamics of change in the sum of the decimal value VSI elements generated by the binary transformations XOR, OR, AND, depending on the number of single elements m of the code sequence of ring code showed in the article.

The gain from the use of the ring code using the vector of shift indices is 2.7 times compared with the amount of information transmitted.

References

1. Tolubko V.B, Otrokh S.I., Berkman L.N., Pliushch O.G., Kravchenko V.I. Noise Immunity Calculation Methodology for Multi-Positional Signal Constellations // 14th IEEE International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET'2018): Conference Proceedings. – Lviv, 20-24 of February 2018. – Paper 436.
2. Otrokh S.I., Hryshchenko L.M., Dubrovsky V.V., Melnik U.V.Peculiarities of the Formation of Commemoration of Kilts Kodiv Type 001011: Mathematical Model – Kyiv: Communication –2018 – № 1 – p.33-40 (in Russian).
3. Tolubko V.B, Otrokh S.I., Berkman L.N., Kravchenko V.I. Manipulation coding of signal n-dimensional multi-position constellations based on the optimal noise immunity of regular structures – Kyiv: Telecommunication and information technology –2017 – № 3(56) – p.5-11 (in Russian).
4. Otrokh S.I., Kuzminykh V.A., Sosnovsky I.O. Future network in action, online life – Kyiv: Communication –2018 – № 6 – p.42-45 (in Ukrainian).
5. Hryshchenko L.M. Patterns of formation of ring codes. Mathematical model – Kyiv: Communication – 2016 – № 5(123). – p.27-31 (in Ukrainian).
6. Hryshchenko L.M. Mathematical model for creating the 010101 type family ring code – Kyiv: Communication – 2017– № 1(125). – p.58-61 (in Ukrainian).
7. Havrylko E.V., Otrokh S.I., Yarosh V.I., Hryshchenko L.M. Improving the quality of the future network by using the ring – Minsk: Communication Herald –2018 –№2 –p.60-64 (in Russian).

APPLICATION OF DECISION-MAKING METHODS FOR EVALUATION OF COMPLEX INFORMATION SYSTEM FUNCTIONING QUALITY

Hnatiienko H.M.[1], Snytyuk V.Y.[2] and Suprun O.O.[3]
Intellectual and Information Systems Department, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine,
[1]g.gna5@ukr.net, [2] snytyuk@gmail.com, [3] oleh.o.suprun@gmail.com

This article presents an approach, which allows to evaluate and, hence, to improve the functioning of complex intellectual systems. Although most of the intellectual systems have one goal – to create a trustful and accurate simulation of real-life program or event, that will allow to make the best tactical and strategic decisions in long or shirt terms, it's impossible to design universal scheme for these systems, since they vary very much, same as real-life practical tasks. Defining the system quality level gives an expert the opportunity to adequately perceive the results of the simulation. To solve this problem, some heuristics are introduced in the article, to link abstract concept of the model with real-life problems. The proposed method allows to calculate and to compare risks that appear in every company, like those, connected to lack of recourses, or inefficient work of its employees. Besides, according to quality levels, and expert can make decisions about replacing inefficient links of the system, to transfer responsibilities between different links, or rank the tasks inside the system according to their importance for the system as a whole. The proposed method can be used with different systems and environments, since it gives an expert the possibility to set necessary coefficients during the preparation stage.

Keywords: *Complex Information Systems, Decision-Making, Quality Evaluation.*

1 Introduction

The problem of ensuring the quality of information system functioning is especially relevant today. This is due to the need of providing reliable and adequate information right in time to achieve the main tasks of the system functioning in conditions of strict competition in all spheres of human life.

Even the seemingly small mistake, made at the modeling and designing stage, can provide great loss in future. This is especially important for big companies that can be described as complex systems with multiply subsystems, which have different connections between each

other. Loosing one of these connections or subsystems may cause serious damage for the system as a whole, so the risks must be calculated and the most important nods stated in advance. The intellectual systems simulation is the beat way to make a trustworthy model that will react and change itself same as the real object. And the quality evaluation is same important as modeling itself.

To explore the systems functioning, theoretically-gaming, probabilistic, graph and matrix models are traditionally used [1]. To evaluate the quality of the complex information system functioning, the methods of collective objects arrangement, which are a wide class of methods for the simulation of practical problems in various subject areas, will be used [2]. Among the decision-making problems, the task of objects organizing is highlighted by a large number of specific real-life applications and the undoubted topic actuality. The problem of searching for the resulting objects arrangement by individual object arrangement is one of the most common problems of linear objects arrangement.

A complex information system contains hundreds of elements that perform thousands of tasks, and it may have different nature and specification: for example, a map of organization business processes algorithms of a certain hierarchical system interaction, etc. As the companies scale grows rapidly, more and more information is needed to simulate it's work [3]. To calculate and evaluate all this information, different approaches are used, such as neural systems [4] or evolutional technologies [5]. All this is made to calculate the possible risks and to avoid them [6], that is impossible using an intellectual system with low quality.

2 Problem Setteing

Let some resultant (aggregated, collective) arrangement be given as n problems $R^* = (a_{i_1}, ..., a_{i_n})$, $i_j \in I = \{1, ..., n\}$, $j \in I$, which is built on logic, that characterizes the processes of some information system functioning. Arrangement R^* is built on the basis of individual ordering tasks that are performed by k elements of the system $R^i = (a_{1_i}, ..., a_{n_i})$, $i \in J = \{1, ..., k\}$, where $n_i, i \in J, -$ the number of tasks in individual arrangements, that are performed by $i-$th elements of the system. Let the $A^i, i \in J$, be the subset of tasks, performed by $j-$th element of the system.

Since $R*$ reflects the logic of solving a collective problem, tasks in individual ranking can have indices that do not coincide with the natural series. For example, tasks $a_3 \succ a_5 \succ a_7$ belong to the ranking R^1, but tasks $a_1 \succ a_4 \succ a_2 \succ a_6$ belong to the ranking R^2. Tasks order indicates the sequence of tasks implementing during the system operating.

At the same time, the tasks performed by the elements are not duplicated, ie $n = \sum_{i \in J} n_i$ – each task in the system is unique and each task in the ranking $R*$ occurs only once: $A^{i_1} \cap A^{i_2}, i_1, i_2 \in J$.

Each task from the set of tasks $A = \{a_1, ...a_n\}$ is characterized by two parameters:

$c_i^0 -$ the nominal price of the execution or the need for the resources, $i \in I$;

$t_i^0 -$ the nominal execution time, $i \in I$.

During the performance of $i-$th task by $j-$th element of the system, the following is known:

$c_i^j -$ the real price of the task, $i \in I, \ j \in J$;

$t_i^j -$ the real time of the task execution, $i \in I, \ j \in J$.

Each element of the system in its regular mode executes its tasks and has limited ability to perform all subset of its tasks. These limits are

$$\sum_{a_i^j \in A^j} c_i^j = C^j, \ \ j \in J, \qquad (1)$$

$$\sum_{a_i^j \in A^j} t_i^j = T, \text{ для } \forall j, j \in J. \qquad (2)$$

Restriction (1) is the cost of tasks performing by an element of the system - an employee's salary analogue in the business processes simulation, and restrictions (2) is the time limitation - analogue to the monthly norms of the working time duration during the organizations functioning, although in general, the time constraints may be different.

During the performance of normative tasks, determined by the nominal tactical and technical characteristics of the system, the requirements of the system and its elements in resources (1) - (2) are constant, and the quality of tasks execution by all subsystems and the system as a whole is 100 percent.

The nominal tactical and technical characteristics of the system are the resources requirements:

$\sum_{i=1}^{n} c_i^{0i} = C^0$ – system execution budget,

$\sum_{i=1}^{n} t_i^{0i} = T^0$ – total time requirement to perform the system functioning.

Since tasks are not duplicated, there is no need for direct redundancy. The redundancy is potential, hidden: the functional moves to another element of the system, when an element, that should perform the task according to the norm, can not do this. But this is due to the additional costs of a limited resource.

It is necessary to design a model that will reflect the system's response to various types of environmental influences and changes in the states of system elements. In this case, the quality of the system and its elements functioning should be evaluated, depending on the system elements state.

3 Modeling the Decision-making Situations

In the process of system performance in real conditions, the situation described in the problem setting, can significantly differ from the normative one. For example, in the case of a large organization, there are always employees who are currently on sick leave, on vacation, on business trips, absent for unknown reasons, formally issued refuses, dismissed from work for various reasons, violate labour discipline etc.

All these reasons can be estimated, heuristically determine the current level of performance for each task and evaluate the quality of each task at the 100 percent scale.

In case of temporary or long-term failure of the system element, all functions that are performed by this element can not be executed by the system. For their implementation, it is necessary to make decisions about the functions redistribution or their replacement. For example, during the temporary absence of the system element, its tasks can be:

– distributed to perform among other elements of the system;

– passed to perform to one element of the system;

– ignored as such, without which the information system will not significantly lose its functionality level.

3.1 Model 1. Tasks distribution among the elements of the system.

It is clear that the tasks distribution can only be done between the elements that can perform these tasks, according to their qualifications,

the available certificates, etc. In this case, such heuristics should be taken into account.

Heuristic H1. While solving tasks that are not normative for the current system element, the quality of these tasks performance by current elements that are intended for the temporary execution is clearly reduced. The level of the tasks performance quality is set individually for each case and can be, for example, 80%.

Heuristic H2. During the necessity of performing by system element of additional tasks, the situation of element overload occurs, and therefore the quality is reduced:

a) performance of its own normative tasks, for example, to the level of 90-95%;

b) performance if additional tasks based on Heuristics 1.

Heuristic H3. The price of resource type (1) in case of tasks redistribution due to the lack of one of the system elements, may increase in the range from 10% to 15% - to increase the motivation of new elements to perform additional tasks.

After applying heuristics H1-H3 the recount of the resources that are needed to perform tasks in new circumstances must be made. It is clear that the new values will differ significantly from the normative ones. At the same time, the quality of the tasks, and, therefore, and the quality of the system will vary greatly from the ideal 100%.

3.2 Model 2. Redistribution of the missing element tasks to another for their execution.

With a significant additional load on the system element, that is assigned to perform the task of a missing element, the quality of the new tasks implementation, and also the tasks that it has been normatively performed, is greatly reduced. In this model the additional heuristic must be used.

Heuristic H4. With an additional load on the system element, the quality of its additional tasks implementation significantly decreases, for example, according to a linear function, the parameters of which can be assigned separately for each situation of decision-making.

Heuristic H5. The load on the system elements can not exceed some given value, for example, 2 * T, where T - time constraints established by the formula (2).

During the application of heuristics H4-H5, the definition of new quality levels for the tasks performance and the quality of the system as a whole is made. In addition, there are changes in the requirements of

resources that are necessary for the system to perform tasks in the new environment – taking into account the transfer of all tasks of the missing element to another element.

3.3 Model 3. Ignoring the tasks that are performed by the missing element of the system.

If it is known that a system element is temporarily absent, and an experienced expert, that makes decisions, understands that there is no urgent need for the task of this element to be performed, a temporary moratorium to perform the relevant tasks may be made.

Heuristic H6. If the element responsible for performing an autonomous task is absent, the quality of the task performance drops gradually, during some time. The regularity of the task quality reduction can be set separately for each individual case.

Heuristic H7. If a task for which an executor is not absent is not autonomous, that is, other tasks depend on its implementation, the function of changing the implementation quality of the dependent tasks is given separately for each specific decision-making situation.

Decision to ignore tasks, that are temporarily left without an executor, is very responsible and requires constant monitoring by the expert or controller appointed by him. At each monitoring iteration, an assessment of the quality changes must be made according to the heuristics H6-H7.

4 The Results of Information System Functioning Quality Evaluation

After making decisions about the functions redistribution between system elements or their replacement, new values of resources for the system tasks and the level of functioning quality are calculated. This information is recorded in the system database.

Based on the obtained values, the affiliation of the system functioning quality levels to the fuzzy set (0,1) is determined. Approaches to the determination of affiliation functions and algorithms for constructing affiliation functions on the basis of the values frequency analysis are given in the monograph [2]. That is, the system functioning quality as a result of the described procedure application will be characterized by the function of affiliation to the fuzzy set.

It is also possible to design functions for a priori introduced linguistic variables with such names as, for example, "critically acceptable quality

of functioning", "risk system operation", "sufficient quality", "excellent quality", etc.

5 Knowledge Base for the Functioning Quality Evaluation

The practical significance of the proposed models will greatly increase if the expert has the tools for evaluating various decision options to ensure that tasks that should be performed by missing elements of the system are performed. To use the described models for information system quality evaluation, it is necessary to create a knowledge base with such indicative content:

– interchangeable elements of the system and the degree of their interchange ability;

– the restriction related to the ability to perform or delegate the tasks performance associated with hierarchical links in the system;

– tasks decomposition in system elements and potential assignment of tasks for critical elements;

– functions of changing the system elements working capacity at non-normative overloads;

– information about the possibility of some tasks duplication by individual elements of the system;

– the priority of tasks implication by several elements - whenever possible and necessary;

– the possibility of a temporary moratorium on some tasks;

– formula for calculating the load for system elements;

– the inclusion of tasks in processes, the criticality of the certain tasks performance, the estimation of the system quality level loss;

– an evaluation of the decrease in the functioning quality in the absence of coordination from the elements that control the hierarchical system;

– taking into account the factors of system quality decreasing: lack of competence of the element that temporarily performs the task, or overloading the element with additional tasks.

6 Possibilities of Applying the Different Models Classes to the Evaluation of the Information System Functioning Quality

In the first stage of modeling, system elements may correspond to a non-oriented graph - only the presence of tasks is indicated, without a detailed description of inputs and outputs.

For systems that perform tasks, where the order of execution is essential, it is necessary to apply models of strict tasks ranking, described in this paper.

If the parallel processes of task execution are modeled, models of non-strict ranking can be applied - for grater detailing.

When there are cycles in the interaction between tasks, it is necessary to apply individual matrices of the tasks sequence - in such cases, the resulting matrix of pairwise tasks ranking will be block-diagonal and substantially sparse.

The metric matrix of pairwise tasks ranking is used in cases where it is essential to specify the terms between the events occurrence or the tasks beginning - for example, when describing the Gantt chart using matrices.

If these terms of tasks are not clear, then matrices of pairwise tasks ranking with elements in the form of affiliation functions can be used to simulate such systems.

7 Perspective for Improving the Adequacy of Modeling the System Quality

For a more complete consideration of the real systems features, it is necessary to complicate the described mathematical model. In particular, this can be done by taking into account the following factors:

– ranking system of the elements, definition of subordination between elements;

– the establishment of hierarchical links between the elements of the system and the determination of the influence levels of one element to another or the absence of such influence;

– definition of a priori priority of tasks, regardless of their importance in terms of cost or execution time;

– taking into account the competence coefficients of system elements;

– increasing the level of detail and the level of adequacy of the model by describing subtasks;

– description of processes that establish the links between tasks and subtasks.

An example of a process consisting of tasks and subtasks that are related to the logic of the system's performance, its organizational and functional structure as a whole, is the following:

Process	System Elements	Tasks	Subtasks
Process Name	1	1	1.1 1.2
	2	2	2.1 2.2 2.3
	1	3	1.3
	3	4	3.1 3.2
	1	5	1.4 1.5

8 Further Research Directions

Based on the described approach, new tasks can be developed and new approaches to increase the adequacy of the modeling can defined:

– a priori assessment of the system reliability;

– evaluation of permissible decrease of the system elements functioning quality and the level of tasks performance;

– considering the presence or absence of links between tasks: the impact of the task on the quality of the other tasks functioning;

– solution of optimization problems of forecasting the system functioning quality, the cost of providing this quality and calculation of allowable time expenditures;

– restoration of the acceptable quality level in case of several elements failure: determination of functioning necessary conditions.

9 Conclusions

In the article the problem setting and different models for system functioning quality evaluation are proposed. Since the real-life problems may be very different, it's very important to establish common problem setting. The variety of proposed models allows an expert to choose one according to the occurred situation without the necessity to design a new model as a whole.

For the problem solving different methods of decision-making theory, including the classic ones, or newly-established approaches, such as artificial intelligence, may be used. This gives the opportunity to choose the best method, according to the problem, and the ability to compare them at the same time. The proposed heuristics display the whole range of real situations, or at least the main ones.

As the flow of this method, the requirements of long-term preparations can be stated, such as choosing the priorities, setting the coefficients or determining the main nodes of the system. At the same

time, this allows to configure the system according to the real-life problem, and, as result, more accurate answer may be obtained.

The perspectives for the future development of such problems are mentioned, same as possible ways of improvements, using new methods, to enhance models matching with real objects

References

1. A.G. Dodonov, D.V. Lande: Vitality of information systems. - K . Science. Dumka, 2011. - 256 p.
2. H. M. Gnatyenko, V. Ye. Snytyuk: Expert decision making technologies. - K.: McLaugh, 2008. - 444 c.
3. Tsan-Ming Choi ; Hing Kai Chan ; Xiaohang Yue: Recent Development in Big Data Analytics for Business Operations and Risk Management, IEEE Transactions on Cybernetics (Volume: 47 , Issue: 1 , Jan. 2017).
4. Miyuan Shan ; Wenxi Zhu: Risk Analysis for Highway Project Based on Multiple-Dimension Neural Network, 2006 6th World Congress on Intelligent Control and Automation.
5. V.Y. Snytyuk ; O.O. Suprun: Evolutionary techniques for complex objects clustering, 2017 IEEE 4th International Conference Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD)
6. Alexander Setiawan ; Adi Wibowo ; Andrew Hartanto Susilo: Risk analysis on the development of a business continuity plan, 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)

EVOLUTION OF CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE IN IMAGE CLASSIFICATION PROBLEMS

Andrey Arsenov[1], Igor Ruban[1], Kyrylo Smelyakov[1], Anastasiya Chupryna[1]

[1] Kharkiv National University of Radio Electronics, Nauky Ave. 14, Kharkiv, 61166, Ukraine
andrii.arsienov@nure.ua, ruban_i@ukr.net, kirillsmelyakov@gmail.com, anastasiya.chupryna@nure.ua

At present, the models and computer vision algorithms are increasingly used in various fields of activity. For example, in systems of sample analysis in medicine and pharmacology, in identification of individuals by a fingerprint, iris or face, in video surveillance security systems and in many other systems and applications. In connection with the growth of computing power and the emergence of big databases of images, it became possible to learn and use deep neural networks for solving the problems of classification and recognition. As to the image classification problem, the Convolutional Neural Networks showed themselves best of all; every year since 2012, they won the prestigious international contest – the ImageNet Large Scale Visual Classification Challenge (ILSVRC), in which such giants as Google and Microsoft participated. Thanks to the revealing of their capabilities, the convolutional neural networks are increasingly used for pattern recognition, image classification, object detection, semantic segmentation, and solving many other problems. The paper examines the evolution of the most efficient models and trends in development of architecture of convolutional neural networks, which are currently used for classification of images that have been included in the list of winners of this international competition, ILSVRC. More precisely, the key features of architecture and its annual variations are revealed on the background of increasing efficiency of practical application of these networks. The data of numerous experiments conducted over the past few years are summarized, classes of applied problems are analyzed, and estimates are given for an effectiveness of use of

the considered convolutional neural networks. In fact, these performance estimates are based on evaluation of probability of adequate classification of images. On this basis, a generalized algorithm is formulated, and practical recommendations are proposed taking into account the problem features.

Keywords: *Convolutional Neural Network, Image Classification, Neural Network Architecture, Efficiency.*

1 Introduction

Computer vision technologies are becoming increasingly popular. They are used in data analysis systems in medicine and pharmacology, in personal identification tasks, by face, by fingerprint, by iris, in security video surveillance systems, for example, for identifying vehicles by their license plates and in many other systems and applications [1-3]. In connection with the growth of computing power and the emergence of huge image bases, it became possible to train deep neural networks for solving problems in the field of computer vision, such as classification and recognition. Convolutional Neural Networks showed themselves best in the image classification task [4-5], which since 2012 each year won the competition of the international competition ImageNet Large Scale Visual Classification Challenge (ILSVRC), in which such giants took part and Microsoft.

A convolutional neural network is a neural network with a convolutional layer. Usually in the convolutional neural networks there are also a sub-sampling layer (pooling layer) and a fully connected layer. Convolutional neural networks are used for pattern recognition, object detection, image classification, semantic segmentation, and other tasks. In convolutional neural networks, layers of convolution and subsampling consist of several “levels” of neurons, called feature maps, or channels. Each neuron of this layer is connected to a small section of the previous layer, called a receptive field. In the case of an image, a feature map is a two-dimensional array of neurons, or simply a matrix. Other measurements can be used if another kind of data is taken as input, for example, audio data (one-dimensional array) or volume data (three-dimensional array) [6-7].

At the same time, although such networks are used quite successfully, the question of choosing the optimal architecture and setting the

parameters of the neural network are remains unresolved. In this regard, the task of the work is to analyze the available experimental data using the most efficient convolutional neural networks used to classify images, in order to develop a general algorithm and practical recommendations on choosing the best architecture and setting the parameters of the neural network, according to the specifics of the problem.

2 The effectiveness of the use of the convolutional neural network for image classification

This section presents the most efficient and widely used architectures of convolutional neural networks for classifying images that are arranged in chronological order.

2.1 Convolutional neural network AlexNet

The first neural network that won the ILSVRC image classification competition was AlexNet, in 2012, reaching a top-5 classification error of 15.31%. For comparison, the method that does not use convolutional neural networks received a classification error of 26.1%. AlexNet collected the latest technology at the time to improve the network. The architecture of this network is shown in Fig. 1.

Training network AlexNet due to the large number of network parameters occurred on two graphics processors (abbreviated GPU – Graphics Processing Unit), which reduced training time in comparison with learning based on the central processor (abbreviated CPU – Central Processing Unit). It also turned out that using the Rectified Linear Unit (ReLU) activation function instead of more traditional functions (sigmoids and hyperbolic tangent) made it possible to reduce the number of learning epochs by six times. This is due to the fact that the function of network activation Rectified Linear Unit allows you to overcome the problem of gradient attenuation inherent in other activation functions. Graphically, the activation function of the Rectified Linear Unit is shown in Fig. 2.

Also, a dropout technique (Dropout) was used in AlexNet, which randomly turns off each neuron on a given layer with a probability p at each epoch. Then, after learning the network, at the recognition stage, the weights of the layers to which the dropout was applied should be multiplied by 1/p. Technology Dropout acts as a regularizer, not allowing

the network to retrain. To understand the effectiveness of this technique, there are several interpretations. First, this dropout causes neurons not to rely on neighboring neurons, but to learn to recognize more persistent signs. And the second, later, is that learning a network with a dropout is an approximation of learning a network of ensembles, each of which represents a network without some neurons. As a result, the probability of error is reduced, since the final decision is made not by one network, but by an ensemble, each network of which is trained differently.

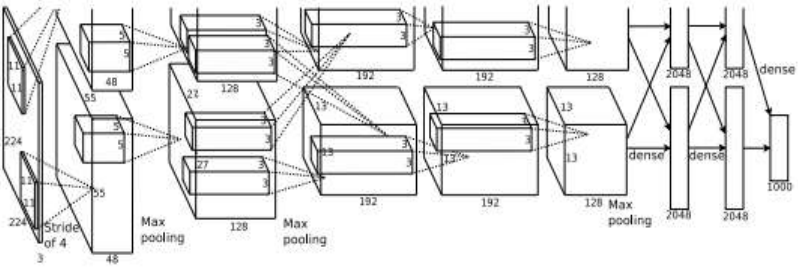


Fig. 1. Architecture of convolutional neural network AlexNet [8].

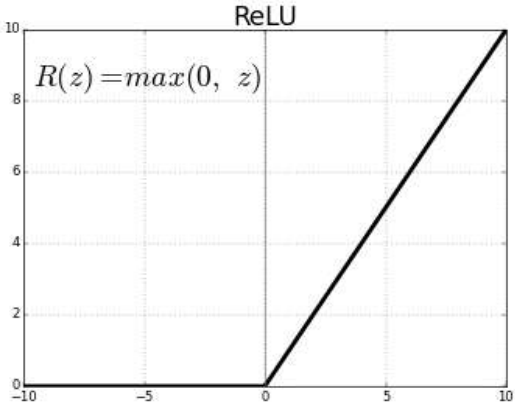


Fig. 2. Network activation function Rectified Linear Unit [9].

2.2 Convolutional neural network ZF Net

The convolutional neural network ZF Net is the winner of ILSVRC 2013 with a top-5 classification error of 14.8%. The main achievement of

this architecture is the creation of a filter visualization technique - a sweep network (deconvolutional network), consisting of operations, in a sense, reverse operations of the network. As a result, the network sweep displays a hidden layer of the network on the original image.

To study the behavior of the filter on a particular image using a trained neural network, you must first make a network output, then in the layer of the studied filter zero all weights, except the weights of the filter itself, and then apply the resulting activation to the network of the sweep network. The network sweeps consistently used operations Unpooling ReLU and filtering. The Unpooling operation partially restores the input of the corresponding sub-sampling layer by remembering the coordinates that the sub-sampling layer has selected. The ReLU operation is a regular layer that uses the ReLU function. The filtering layer performs the convolution operation with the weights of the corresponding convolution layer, but the weights of each filter are "inverted" vertically and horizontally. Thus, the initial activation of the filter moves in the opposite direction until it is displayed in the original image space. The architecture of the considered network is shown in Fig. 3.

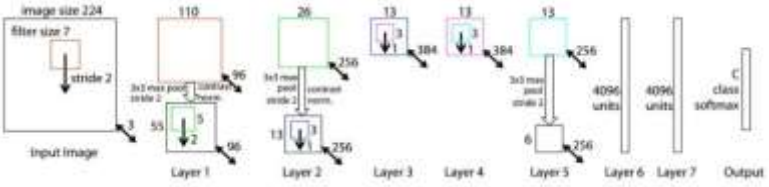


Fig. 3. Network activation function ZF Net [10].

2.3 Convolutional neural network VGG Net

VGG Net is a convolutional neural network model that won the 2014 image classification competition. In this network, they refused to use filters larger than 3x3. Since the authors proved that the 7x7 filter layer is equivalent to three layers with 3x3 filters, and in this case 55% less parameters are used. Similarly, a 5x5 filter layer is equivalent to two layers with a 3x3 filter, which saves 22% of network parameters.

Features of the architecture and internal organization of this neural network are shown in Fig. 4.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Number of parameters (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

Fig. 4. Different variations of the convolutional neural network architecture VGG Net [11].

2.4 Convolutional neural network Inception

The Inveption-v1 convolution neural network is the winner of the ILSVRC 2014 competition with a top-5 error of 6.7%, also known as GoogleNet. The creators of this network, led by Christian Szegedy, proceeded from the fact that after each layer of the network it is necessary to make a choice whether the next layer will be a convolution with a 3x3, 5x5, 1x1 filter or a subsampling layer. Each of these layers is useful – a 1x1 filter reveals a correlation between channels, while larger filters

respond to more global features, and a subsampling layer reduces dimensionality without large loss of information. Instead of choosing which layer should be next, it is proposed to use all layers at once, parallel to each other, and then merge the results into one. To avoid an increase in the number of parameters, a 1x1 convolution is used in front of each convolution layer, which reduces the number of feature maps. Such a block of layers was called an Inception module.

Also, GoogLeNet abandoned the use of a fully connected layer at the end of the network, using the Average Pooling layer instead, which drastically reduced the number of parameters in the network. Thus, GoogLeNet, consisting of more than one hundred basic layers, has almost 12 times fewer parameters than AlexNet (about 7 million parameters against 138 million).

In the next iteration of the Inception module, called Inception-v2, the authors, as was done on the VGG network, decomposed the 5x5 layer into two 3x3 layers. Next, the Batch Normalization technique was used, which allows to multiply the learning speed by means of normalizing the distribution of layer outputs within the network.

In the same article [12], the authors proposed Inception-v3. In this model, they developed the idea of filter decomposition, proposing to decompose the NxN filter with two successive 1xN and Nx1 filters. Also in Inception-v3, RMSProp is used instead of the standard gradient descent and truncated gradients are used to increase the learning stability. An ensemble of four Inception-v3 received a top-5 error of 3.58% at ILSVRC 2015, losing to ResNet.

2.5 Convolutional neural network ResNet

The winner of the ILSVRC 2015 competition with a top-5 error of 3.57% was an ensemble of six networks of the ResNet (Residual Network) type, developed at Microsoft Research. The authors of ResNet have noticed that with the addition of new layers, the quality of the model grows to a certain limit (see VGG-19), and then begins to fall. This problem is called the degradation problem, a decrease in accuracy on the validation set.

The authors were able to find such a topology in which the quality of the model grows with the addition of new layers. A neural network can approximate almost any function, for example, some complex function H (x). Then it is true that such a network will easily learn the residual

function: F (x) = H (x) – x. Obviously, that our initial objective function will be H (x) = F (x) + x. If we take a certain network, for example, VGG-19, and add twenty layers to it, we would like the deep network to behave at least as good as its shallow analogue.

The problem of degradation implies that a complex nonlinear function F (x), obtained by adding several layers, must learn the same transformation, if the previous layers had reached the quality limit. But this does not happen; it is possible that the optimizer simply cannot cope with adjusting the weights so that a complex non-linear hierarchical model does the same transformation. In order to "help" the network, it was proposed to introduce a missing connection (Shortcut Connections). The architecture features of this neural network is shown in Fig. 5.

2.6 Convolutional neural networks Inception-v4 and Inception-ResNet

After the success of applying the ResNet convolutional neural network, the following versions of the Inception network were introduced: Inception-v4 and Inception-ResNet. In both cases, the Inception module was divided into modules A, B, and C for inputs with dimensions of 35x35, 17x17, and 8x8, respectively. Reduction blocks were also identified, in which the dimensionality decreases and the depth of the data inside the network increases. In Inception-v4, the main innovations are the replacement of Max Pooling with Average Pooling in the Inception modules themselves.

For Inception-ResNet, skipping connections have been added to the Inception modules. Two versions of the network were designed – Inception-ResNet-v1, which requires less computation, and Inception-ResNet-v2.

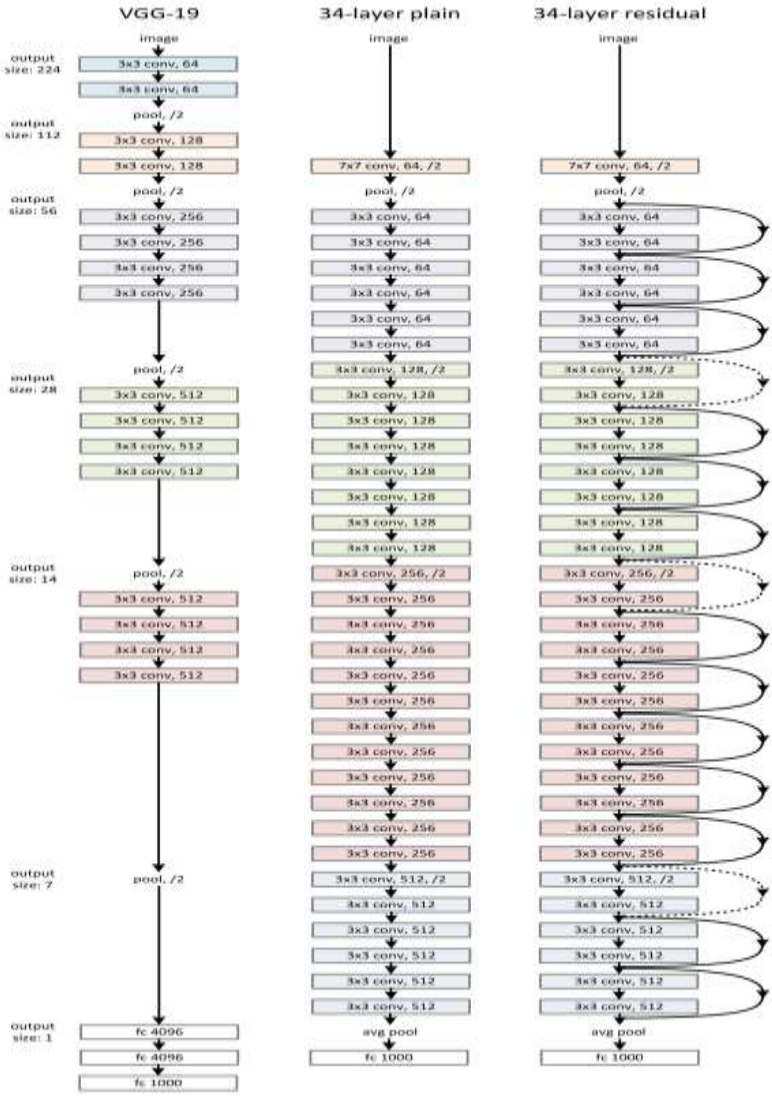


Fig. 5. The architecture of the convolutional neural network Residual Network [13].

2.7 Obtaining of estimates and analysis of effectiveness using of the considered models of convolutional neural networks

To estimate the convolutional neural network models in addition to the type of errors usually indicate the number of models in the ensemble and the number of notches images that were fed to the input of each model. For example, 10 notches means that four notches are made at the corners of the image, one notch in the center, and each notch is additionally horizontally inverted.

According to numerous experiments [10-14], the generalization and analysis of the obtained results were made.

In the Tab. 1 shown the results of the considered neural networks with one model and one cutout based on ImageNet images (except ResNet-152, for which the result for 10 notches is indicated).

Table 1. Network efficiency for a single cut-out model.

Neural network	Top-1	Top-5	Number of layers	Number of operations (G-
AlexNet	39,7 %	18,9 %	8	70 M
ZF Net	37,50 %	14,8 %	8	70 M
VGG Net	25,60 %	8,10 %	19	155 M
GoogLeNet	29,00 %	9,20 %	22	10 M
Inception-v3	21,20 %	5,60 %	101	35 M
Inception-v4	20,00 %	5 %	152	35 M
Inception-ResNet-v2	19,90 %	4,90 %	467	65 M
ResNet-152	19,38 %	4,49 %	152	65 M

In Tab. 2 shown the results of using ensembles of models with many cutouts based on ImageNet images.

As can be seen from these tables, for five years, from 2012 to 2016, the Top-5 error on ImageNet for single models decreased almost four times (from 17% to 4.49%), and for the ensemble – almost five times (from 15.40% to 3.10%).

Analyzing the experimental data (Tab. 1, Tab. 2), we can conclude that the choice of network architecture is made according to the following criteria: classification errors, performance, and the complexity of learning a neural network. For this, the following algorithm is usually used.

Initially, guided by certain requirements, they set a permissible classification error. For example, it is currently believed that a classification error when using human vision is in the range from 5% to 10%. If you look at the classification error of the latest convolutional neural networks, you can see that they are coping with this task as well as a human. This means that in the classification problems that a person solved classically, you can choose any network with an error not higher than the specified one. Based on the analysis of data in Tab. 2, we can conclude that the last five networks will suit us. But we need one. What should be done?

Table 2. The effectiveness of the network for ensembles with many notches.

Neural network	**Models**	**Notches**	**Top-1**	**Top-5**
AlexNet	7	1	36,70 %	15,31 %
ZF Net	6	10	36 %	14,70 %
VGG Net	2	150	23,70 %	6,80 %
GoogLeNet	7	144	—	6,67 %
Inception-v3	4	144	17,20 %	3,58 %
ResNet-152	6	144	—	3,57 %
Inception-v4 + 3x Inception-ResNet	4	144	16,50 %	3,10 %

Next, the selection of an admissible network is made in order to satisfy the specified restrictions on labor intensity (estimates of labor intensity are given in Tab. 1), taking into account the available hardware capacities. This choice is also made taking into account the time constraints on the network learning process, since with the increase in the number of layers and network parameters, the training time will also increase.

The choice of complexity can also be ambiguous, since at the first stage five networks were chosen. In such a situation, the most important criterion is usually identified and the best network is selected by this criterion.

To improve the quality of the classification results, it is planned to use specialized frame preprocessing models and algorithms [15-19] in addition to developing of network ensembles.

3 Conclusion

In the course of considering the most effective models of convolutional neural networks used in our time for the purposes of image classification, an analysis of their architectural features was performed. According to numerous experiments, a generalization and analysis of the results of the efficiency of using neural networks for image classification (Tab. 1, Tab. 2) was made. On this basis, a generalized algorithm is formulated and practical recommendations are given regarding the choice of the best architecture of a neural network, respectively, the specifics of the problem.

References

1. Rafael C. Gonzalez, Richard E. Woods Digital Image Processing, 4th edition Pearson/Prentice Hall, 2018. – 1168p.
2. David A. Forsyth, Jean Ponce Computer Vision: A Modern Approach (2nd ed.). – Pearson Education Limited, 2015. – 792p.
3. Milan Sonka, Vaclav Hlavac, Roger Boyle, Image Processing, Analysis, and Machine Vision (4th ed.). – Cengage Learning, 2014. – 896p.
4. Ian Goodfellow, Yoshua Bengio, Aaron Courville Deep Learning. – MIT Press, 2016. – 787p.
5. Peter Norvig , Stuart Russell Artificial Intelligence: A Modern Approach, Global Edition. – Pearson Education Limited, 2016. – 1152p.
6. Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. International journal of computer vision 88(2), pp. 303-338.
7. Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik Rich feature hierarchies for accurate object detection and semantic segmentation. The IEEE Conference on Computer Vision and Pattern Recognition. – 2014. – pp. 580-587.
8. Papers, https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

9. Medium, https://medium.com/@kanchansarkar/relu-not-a-differentiable-function-why-used-in-gradient-based-optimization-7fef3a4cecec.
10. Arxiv, https://arxiv.org/pdf/1311.2901v3.pdf.
11. Arxiv, https://arxiv.org/pdf/1409.1556v6.pdf.
12. Cv-foundation, https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Szegedy_Going_Deeper_With_2015_CVPR_paper.pdf.
13. Arxiv, https://arxiv.org/pdf/1512.03385v1.pdf.
14. Ross Girshick Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision. – 2015. – pp. 1440-1448.
15. Ruban, K. Smelyakov,V Martovytskyi, D. Pribylnov and N. Bolohova Method of neural network recognition of ground-based air objects // IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT), 24-27 May 2018. – P. 589-592. DOI: 10.1109/DESSERT.2018.8409200
16. K. Smelyakov, D. Pribylnov, V. Martovytskyi, A. Chupryna Investigation of network infrastructure control parameters for effective intellectual analysis // IEEE 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET), 20-24 Feb. 2018. – P. 983-986. DOI: 10.1109/TCSET.2018.8336359
17. K. Smelyakov, A. Chupryna, D. Yeremenko, A. Sakhon, V. Polezhai Braille Character Recognition Based on Neural Networks // IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 21-25 August 2018. – P. 509-513.
18. S. Mashtalir, O. Mikhnova, M. Stolbovyi Sequence Matching for Content-Based Video Retrieval// IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 21-25 August 2018. – P. 549-553.
19. G. Churyumov, V. Tokarev, V. Tkachov and S. Partyka, "Scenario of Interaction of the Mobile Technical Objects in the Process of Transmission of Data Streams in Conditions of Impacting the Powerful Electromagnetic Field", 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 2018. – DOI: 10.1109/DSMP.2018.8478539.

ВИКОРИСТАННЯ ОНТОЛОГІЙ ДЛЯ АНАЛІЗУ МЕТАОПИСІВ У BIG DATA

Гладун А.Я.[1], Рогушина Ю.В.[2], Прийма С.М.[3]

[1] ***Інститут спеціального зв'язку та захисту інформації, НТУ-КПІ ім. Сікорського, Київ, Україна***
Міжнародний науковий центр інформаційних технологій і систем НАНУ, Київ, Україна
glanat@yahoo.com
[2] ***Інститут програмних систем НАН України, Київ, Україна***
ladamandraka2010@gmail.com
[3] ***Таврійський державний агротехнологічний університет, Мелітополь, Україна***
Інститут програмних систем НАН України, Київ, Україна
pryima.serhii@gmail.com

Великі дані (Big Data) — набори інформації (як структурованої, так і неструктурованої або слабоструктурованої) настільки великих розмірів, що традиційні способи та підходи (засновані на рішеннях класу бізнес-аналітики та системах керування базами даних) не можуть бути застосовані до них. Для Big Data характерне феноменальне прискорення нагромадження даних та їх ускладнення. Дуже часто під Big Data у різних контекстах мають на увазі як дані великого об'єму, так і набір інструментів та методів обробки Big Data.

Набори Big Data супроводжуються метаданими, які несуть у собі велику кількість інформації про дані, у тому числі і значну описову текстову інформацію, розуміння машинами якої привело до отримання кращих результатів вирішення поставленої задачі на основі оброблення Big Data. Для підвищення ефективності усіх етапів обробки Big Data з успіхом почали застосовувати методи штучного інтелекту та інтелектуальних Web-технологій до обробки. Найбільш часто така інтеграція стосується використання машинного навчання для здобуття знань з Big Data та онтологічного аналізу – для застосування знань предметної області до аналізу Big Data.

У роботі автори представили метод аналізу метаданих Big Data, який дозволяє відбирати серед гетерогенних джерел та

сховищ великих даних ті блоки даних, які придатні для вирішення задачі, поставленої замовником. Значну увагу зосереджено на обробці текстової частини метаданих (анотації метаданих) та тексту, який описує завдання, що потребує вирішення. Для аналізу та співставлення цих природномовних неструктурованих або слабо структурованих текстів запропоновано використовувати методи аналізу природомовних текстів з використанням онтології Big Data, що містить знання про специфіку цієї предметної області та дозволяє обробляти елементи метаопису Big Data. Створення прототипу такої онтології та подання архітектури інтелектуальної системи співставлення анотацій Big Data з використанням тезаурусів також є складовими цієї роботи.

Ключові слова: *Big Data, метадані, онтологія предметної області, тезаурус, природомовний текст, омонімія, мультимедійні дані, стандарт.*

Вступ

Термін Big Data ("великі дані") стосується групи технологій та методів, за допомогою яких аналізують та обробляють дані великого обсягу, які не піддаються обробці традиційними способами, як структурованих, так і неструктурованих, для отримання якісно нових знань. Актуальність цього напрямку IT визначається експоненційним зростанням обсягів даних, що генеруються в електронній формі та зберігаються у сховищах даних для подальшого використання. Аналіз великих наборів даних є міждисциплінарним завданням, що поєднує у собі математику, статистику, комп'ютерні науки й спеціальні знання предметної області.

Для ефективного практичного використання Big Data виникає потреба їх аналізу на семантичному рівні.

1 Визначення Big Data

Певний набір даних доцільно розглядати як Big Data, якщо йому притаманні одна чи кілька наступних характеристик:

• *обсяг (*volume) – великі обсяги потребують спеціалізованих засобів обробки;

• *швидкість (*velocity) – дані можуть накопичуватися з високою швидкістю;

• *різноманіття* (variety) – дані можуть бути представлені у різноманітних форматах і типах даних, що ускладнює їх інтеграцію;

• *достовірність* (veracity) – дані можуть містити помилки та шум, які не можуть бути перетворені в інформацію і, отже, не мають цінності.

• *цінність* (value) – тільки частина даних може бути корисною.

Основними типами Big Data можуть бути: структуровані дані (SQL бази даних); слабоструктуровані дані (інструкції з інформаційної безпеки, дані профілю клієнтів, журнали Web-серверів, Web-сайти, тексти, електронні листи тощо) і неструктуровані (не реляційні, NoSQL бази даних) дані (аудіофайли, відеофайли, зображення, інформаційні куби тощо). Big Data забезпечують об'єднання (зв'язування) територіально розподілених наборів даних, враховуючи такі операції як реплікація та шардинг (розбиття на фрагменти). Крім того, для Big Data характерно об'єднання різноманітних незв'язаних наборів даних, обробка великих обсягів неструктурованих даних (частка яких у загальному обсязі Big Data є найбільшою).

Сьогодні людство генерує усе більші об'єми Big Data, але ця інформація не має цінності безпосередньо, а здобувається тільки в результаті переробки та аналізу даних. Через величезні обсяги та швидкість надходження інформації така обробка може виконуватися тільки автоматизовано. Отримані після обробки знання, можуть мати практичну цінність такого типу:

- правила, побудовані засобами машинного навчання;
- результати застосування цих правил до аналізу нових даних.

До першого типу відносяться, наприклад, дерево рішень для задачі медичної діагностики або багатошарова нейронна мережа, яку навчили ідентифікувати людей за фотографіями в соціальній мережі.

До другого типу відносяться, наприклад, постановка діагнозу конкретному пацієнту на основі дерева рішень або розпізнавання нейронною мережею особи, зображення якої отримав користувач соціальній мережі.

Для здобуття цих знань з Big Data використовують статистичну обробку та машинне навчання (ML) [2]. Не розглядаючи детально методи та можливості ML, слід відмітити, що машинне навчання – це узагальнення певного досвіду системи, збереженого в електронній формі, для подальшого вдосконалення поведінки цієї системи для

більш ефективного виконання своїх функцій. Його результати мають ймовірнісний, статистичний характер, і значною мірою їх якість залежить від того, наскільки дані, що оброблялися, близькі до тих, що використовуються на практиці. Таким чином, дуже актуальною є проблема знаходження саме тих масивів Big Data, що пертинентні конкретній задачі користувача (містять неявно потрібні знання), достовірні, актуальні та якісні. Ці параметри Big Data оцінюють не безпосередньо, а через аналіз їх метаданих.

2 Огляд проблем в Big Data

У [3] визначено головні проблеми, які існують сьогодні в технології Big Data і потребують вирішення:

1. Проблема інтеграції даних, яку можна подати як комбіновану проблему, що потребує: (1) визначення задачі, яку необхідно вирішити за допомогою Big Data; (2) виявлення (пошук) відповідних частин даних в сховищах та джерелах Big Data; (3) виконання ETL у відповідних форматах та збереження даних для подальшої обробки; (4) зняття неоднозначності даних (приміром, омонімії); (5) обробка даних для вирішення задачі.

2. Проблема подолання гетерогенності між різними наборами великих даних. Семантика може розглядатися як засіб для створення моста між гетерогенними даними.

3. Проблема зв’язування відкритих даних (Linked Open Data) для забезпечення хорошого зв’язування даних.

4. Проблема використання семантики для інтеграції даних та в розробці майбутніх СКБД. Більш того, семантика може бути використана в існуючій системі, для виявлення невідповідностей даних, генерування нових знань за допомогою машини логічного виведення або просто зв'язувати більш точно конкретні дані, що не мають відношення до методів машинного навчання.

Як показує аналіз наукових публікацій [4] сьогодні питання про актуальність метаданих, використовуваних у великих даних є найбільш гострим ніж будь-коли. Усе більше і більше організацій усвідомлюють, що для того, щоб підвищити ефективність бізнесу, використовуючи цінну інформацію від даних, необхідні робастні (працездатні) метадані для здобуття необхідного контексту та походження ключових активів даних. У той же час регулювання галузі вимагає кращої прозорості та розуміння інформації метаданих.

Хоч керування метаданими відоме уже десятки років, сьогодні розробляються нові стратегії та підходи:

- Підтримка постійного розвитку середовища упорядкування даних;
- Пошук більш ефективних шляхів керування бізнесом за допомогою метаданих.

Таким чином, нам потрібно виконати огляд стратегій та технологій роботи з метаданими, доступних для сучасної організації, а також з'ясувати питання, як побудувати успішні стратегії стосовно прийняття та використання метаданих.

Організації та компанії зацікавлені в двох типах великих даних. 1. Для оброблення дуже часто застосовують дані створені людиною, як такі, що в основному розповсюджуються через засоби Web (соціальні мережі, файли cookie, електронні листи, онлайн телебачення, онлайн мовлення тощо). 2. Існує потреба в інтеграції даних, що генеруються різними джерелами і часто є гетерогенними за своєю природою. Приміром, технологій Інтернет людей (Internet of Human) та Інтернет речей (Internet of Things) генерують змішаний трафік великих даних, який використовують сумісно для передбачу вального (предикативного) аналізу для отримання знань щодо розуміння, планування та упередження дій для цих систем. При цьому гостро постає питання стосовно якості даних. Насправді, оскільки великі дані характеризуються великими об’ємами, є «сирими» по своїй природі. Тому потрібне вирішення цієї проблеми.

3 Постановка задачі

Необхідно розробити метод аналізу метаданих Big Data, який дозволяє відбирати серед гетерогенних джерел та сховищ великих даних тих блоків даних, придатних для вирішення задачі, поставленої замовником. Слід враховувати, що як постановка задачі, так і анотації Big Data – це природномовні (ПМ) неструктуровані або слабо структуровані тексти. Тому для їх співставлення доцільно використовувати методи аналізу ПМ, але з використанням онтології Big Data, що містить знання про специфіку цієї предметної області (ПрО) і дозволяє обробляти семантично інші елементи метаопису Big Data (співставляти параметри структури метаопису з поняттями ПрО). Створення прототипу такої онтології також є складовою даної роботи.

4 Напрямки інтеграції інтелектуальних Web-технологій з обробкою Big Data

Аналіз публікацій свідчить про високий інтерес до застосування методів штучного інтелекту та інтелектуальних Web-технологій до обробки Big Data. Найбільш часто така інтеграція стосується використання машинного навчання для здобуття знань з Big Data та онтологічного аналізу – для застосування знань ПрО до аналізу Big Data. Наукові дослідження в цій сфері – розробка відповідних моделей і методів та оцінка їх ефективності – є сьогодні одним з пріоритетних напрямків наукових досліджень.

Прикладом інтересу до цього напрямку є Онтологічний самміт 2017 "AI, learning, reasoning and ontologies" [5], на якому способи використання методів ШІ для машинного навчання, логічного виведення і онтологічного аналізу, орієнтовані на застосування до даних великого обсягу, розглядалися в таких напрямках:

- застосування ML для здобуття знань і поліпшення онтологій;
- використання фонових знань для поліпшення результатів ML;
- використання онтологій для логічного виведення і навпаки.

Онтологічний аналіз та логічне виведення в обробці Big Data засобами ML забезпечує використання фонових знань для і підготовки даних для навчання і тестування (скорочення великих, зашумлених наборів даних до керованих) та усунення неоднозначності термінів. Щоб перейти до етапу навчання, потрібно визначити:

- яку задачу вирішує комп'ютерна система;
- в якому напрямку потрібно вдосконалити її поведінку (приміром, підвищити точність розпізнавання, розширити кількість осіб, що ідентифікуються, пришвидшити розпізнавання);
- звідки отримати дані, в яких містяться потрібні для аналізу відомості (з досвіду взаємодії даної системи з конкретним користувачем чи з усією спільнотою користувачів, із зовнішніх джерел, від аналогічних систем тощо);
- як інтегрувати отримані результати з набором знань, якими оперує система.

Якщо виявляється, що доцільно застосовувати зовнішній досвід, що представлений у Big Data, то виникає проблема пошуку відповідних джерел.

Для цього необхідно використовувати метадані, якими супроводжуються Big Data, та аналізувати їх семантику. Та частина метаданих, що генерується автоматично, містить недостатньо відомостей про те, чи можна здобути з них необхідні знання, і тому виникає проблема семантичного аналізу анотацій цих даних.

Такі анотації, що створюються при збереженні Big Data у відповідних сховищах, можна розглядати як неструктуровані або слабко структуровані природномовні (ПМ) тексти й використовувати для них стандартні засоби аналізу ПМ, аналогічні до засобів пошуку в Web. На жаль, у загальному випадку така задача аналізу вирішується не ефективно, і тому доцільно застосовувати додатково апріорні знання щодо Big Data.

Аналіз публікацій показав, що, незважаючи на високий інтерес до Big Data та наявність різноманітних технологічних засобів їх збереження та обробки, на сьогодні відсутні стандарти метаданих, специфічні для Big Data. Це пояснюється складністю та різноманіттям самих Big Data. Наявні метадані – це технічна інформація, яка характеризує час створення контенту, його обсяг, формати тощо, але не стосується змісту тієї інформації, яка міститься в цих даних. Це унеможливлює уніфікований опис їх семантики. Але значна частина Big Data супроводжується певними анотаціями або поясненнями, поданими, як правило, природною мовою. Тому виникає проблема аналізу таких анотації та визначення за ними пертинентності певних масивів Big Data тій задачі, для вирішення якої ці дані будуть аналізуватися.

Іноді потреби в такому співставленні не виникає – приміром, в організації аналізують Big Data, що накопичуються в процесі її функціювання.

Але досить часто виникає ситуація, коли Big Data для аналізу отримуються з різних зовнішніх джерел. Необхідність попередньої фільтрації контенту пов’язана з тим, що задачі аналізу Big Data базуються на методах машинного навчання, швидкість роботи яких залежить від обсягів інформації, що обробляється. Приклад такої задачі – аналіз потоків телебачення та радіомовлення. В такому випадку доцільно не витрачати час на аналіз усієї інформації, а спочатку відібрати ту частину передач, що стосуються певної проблеми. Джерелом анотацій в такому випадку може бути програма телепередач.

Інший приклад Big Data, що отримуються з різних джерел та анотуються лише природомовними описами – інформаційні ресурси про наявність робочих вакансій Європейської служби зайнятості (EURES – European Employment Services), яка об'єднує близько 400 «євро-радників» з національних служб зайнятості, асоціацій роботодавців, профспілок, місцевих та регіональних органів влади і вищих навчальних закладів активно використовує класифікатор ESCO (European Skills, Competences, Qualifications and Occupations, багатомовного класифікатора європейських навичок, умінь, кваліфікації та професій.

5 Метадані для Big Data

Для одержання якісних результатів обчислення великих даних необхідний супровід блоків даних метаданими і фізично вони приєднуються до блоків Big Data. Метадані надають інформацію про характеристики та структуру набору Big Data. Відстеження метаданих має ключове значення при обробці Big Data, а також для їхнього збереження й аналізу, оскільки вони надають інформацію про походження даних, про джерело даних під час обробки [6, 7]. Наприклад, метадані містять у собі: найменування; інформацію про джерело; XML-теги, що вказують автора і дату створення документа; атрибути, що вказують розмір і формат, контрольну суму; кількість записів набору даних; і дозвільна здатність файлу цифрової фотографії; короткий опис даних.

Метадані – це структуровані, кодовані дані, які описують характеристики об'єктів-носіїв інформації, що сприяє ідентифікації, виявленню, оцінці і керуванню цими об'єктами. Метадані необхідні для опису значення і властивостей інформації з метою кращого її розуміння, класифікації, керування і використання даних.

Властивості метаданих, їхній склад і функції істотно залежать від технологій реалізації систем, у яких вони використовуються, особливостей описуваних ними ресурсів, а також від області застосування й конкретних застосунків.

Метаданим присвячено величезну кількість публікацій, проте, трактування терміна "метадані" усе ще не сформувалось остаточно. Метадані є особливим видом інформаційних ресурсів, їх створення часто вимагає значних зусиль і істотних витрат, однак вони істотно підвищують цінність даних, забезпечують більш широкі можливості їхнього використання.

На сьогодні існує багато визначень метаданих. Ми вибрали найбільш істотні з них, наприклад: метадані – це дані про даних [8]; метадані – це інформація, яка робить дані корисними [9]; метадані – це машинно-оброблювані дані, які описують деякі ресурси, цифрові і не цифрові [10]; метаданими називають інформацію, яка припускає її комп'ютерну обробку і інтерпретацію людиною про цифрові та нецифрові об'єкти [11]; метаданими називається структурована інформація, яка описує, пояснює, вказує місцезнаходження і. таким чином, полегшує пошук, використання інформаційного ресурсу, а також керування ним [12]; у Web – це слабоструктуровані дані, як правило, узгоджені з відповідними моделями, що забезпечують операційну інтероперабельність в неоднорідному середовищі [13].

Таким чином, підсумувавши указані вище визначення метаданих, можна зробити висновок стосовно їх ролі у Big Data. Метадані сприяють підвищенню якості даних, яка визначається наступними характеристиками: погодженістю (чи є представлення даних однорідним, чи існують дублікати даних, що перетинаються або конфліктують); повнотою (чи всі дані наявні); точністю (збігом збережених і фактичних значень); своєчасністю (чи є актуальним збережене значення). Також метадані забезпечують покращення аналізу даних (OLAP, OLTP, Data Mining), де вони необхідні для розуміння ПрО джерела даних для застосування адекватного обчислення й інтерпретації результатів. Метадані забезпечують застосування загальної термінології і мови взаємодії усередині компанії чи організації, усувають двозначність і забезпечують погодженість висновків усередині компанії.

Обробка Big Data тісно зв'язана з метаданими, особливо при обробці слабоструктурованих і неструктурованих даних. Важливо відзначити, що кожного разу, коли великі дані змінюють свій стан, вони повинні ініціювати збір інформації про походження, що відразу ж записується як метадані. Коли дані попадають в аналітичне середовище, запис їхнього походження може бути ініціюватися записом інформації, що фіксує життєвий цикл даних. Ціллю одержання походження є можливість аргументування отриманих аналітичних результатів, знаючи походження даних, і аналіз кроків чи алгоритмів, використовуваних для обробки даних, що привели до наявного результату. Інформація про походження має важливе значення для розуміння цінності аналітичного результату. Подібно науковим дослідженням, якщо результати не можна виправдати і повторити, вони не заслуговують довіри.

Таким чином, для ефективної обробки Big Data і одержання цінних знань необхідна гнучка структура керування процесом обробки на основі метаданих, які дозволяють створити універсальне середовище для забезпечення інтероперабельності гетерогенних блоків даних, стандартизувати етапи обробки й еволюціонувати платформи обробки.

6 Стандарти для метаданих, застосовні для Big Data

На етапі відбору Big Data виконується співставлення їх метаописів із описом задачі користувача, тому що через надзвичайно великий обсяг та слабку структурованість порівнювати із описом задачі сам контент Big Data недоцільно.

У стандартах серії ISO/IEC 11179 метадані визначені як дані, які визначають і описують інші дані. Це означає, що метадані є даними, а дані стають метаданими, коли вони використовуються таким чином. Це відбувається за конкретних обставин, для конкретних цілей, з визначеними перспективами, без яких дані не є метаданими. Набір обставин, цілей чи перспектив, для яких деякі дані використовуються як метадані, називають *контекстом*. Таким чином, метадані – це дані про дані у деякому контексті.

Метадані можуть зберігатися в базі даних і бути організованими з використанням якої-небудь моделі. Модель, що описує метадані, називається метамоделлю. Приміром, концептуальна модель, представлена в ISO/IEC 11179-3, є метамоделлю в цьому змісті.

Зважаючи на відсутність специфічних для Big Data стандартів для метаданих, доцільно проаналізувати ті існуючі стандарти метаданих, які використовуються до інформації, що може мати властивості 5V та дозволяють відображати семантику контенту.

Значна частка Big Data – це мультимедійна інформація. Існує багато форматів подання мультимедійної інформації форматів запису файлів, що розроблені різними виробниками програмних продуктів і апаратних засобів, проте на сьогодні немає єдиного стандарту, спільного для всіх, тому що кожен виробник розробляє свій унікальний, зручний для його використання підхід, який згодом може одержати поширення. Існуючі формати збереження мультимедіа в електронній формі (GIF, TIFF, PIC, PCX, JPEG, PNG тощо) відрізняються методами стиснення інформації, видами кодувань, призначенням.

Експертна група Moving Picture Experts Group Об'єднаного Комітету зі Стандартизації запропонувала сімейство стандартів

подання мультимедійної інформації MPEG [14]. MPEG-файли займають значно менше місця порівняно з інформацією в інших поширених форматах. MPEG-1 (ISO/IEC 11172) [15], MPEG-2 (ISO/IEC 13818) [16], MPEG-4 (ISO/IEC 14496) [17] – стандарти стиснення мультимедійної інформації.

MPEG-7 («Multimedia Content Description Interface» "Інтерфейс для опису контента мультимедіа" ISO/IEC) [18] – стандарт, орієнтований на семантичне подання мультимедіа. В своїх описах він припускає різний ступінь деталізації. MPEG-7 містить засоби опису – DT (Description Tools); мову опису визначень DDL (Description Definition Language) та системні засоби. Він визначає стандартний набір дескрипторів для різних типів інформації, стандартизує спосіб визначення своїх дескрипторів і їхнього взаємозв'язку. DT містять дві компоненти: дескриптори визначають синтаксис і семантику кожної властивості (елемента метаданих), а схеми опису встановлюють структуру і семантику відношень між їх компонентами, що можуть бути як дескрипторами, так і схемами опису.

Оскільки описові можливості мають однозначно і повністю інтерпретуватися в контексті застосування, то вони для різних доменів користувачів і різних застосувань є різними, тобто той самий матеріал може бути описаний через різні типи властивостей, що відповідають області застосування і можливостям застосування. Приміром, графічне зображення на найнижчому рівні абстракції може бути описане через форму, розмір, текстуру, кольори, палітру, траєкторію руху та положення; а аудіо через тональність, зміни темпу, положення в звуковому ряд, тоді як на верхньому рівні буде подана семантична інформація *«Це сцена з зеленим автомобілем, який їде дорогою, що знаходиться ліворуч, і людиною в білому, яка переходить дорогу праворуч, у супроводі фонового звуку дощу»*. Можуть існувати також проміжні рівні абстракції. Рівень абстракції пов'язаний зі способом здобуття інформації: багато низькорівневих властивостей можуть бути витягнуті автоматично, тоді як високорівневі властивості вимагають втручання людини.

У багатьох випадках бажано використовувати для опису мультимедійних ресурсів текстову інформацію. Однак проблема полягає в тому, що ці описи мають бути як можна більш незалежними від мови, що використовується. Це особливо важливо при обробці імен авторів, назв, місць тощо. Засоби опису MPEG-7 Description Tools дозволяють створити описи контента (тобто набір схем опису DS та відповідних дескрипторів D), що містять

інформацію про створення та використання контента; дійсність, відображену в контенті; набір об'єктів тощо.

MPEG-21 [19] – стандарт «Multimedia Framework», що призначений для створення інфраструктури керування контентом у розподіленому середовищі для семантичного пошуку. Він визначає основні синтаксис та семантику елементів мультимедіа, залежності між ними та операції, які вони підтримують. Він призначений для встановлення інтероперабельності між мультимедійними інформаційними ресурсами.

Проаналізувавши існуючі засоби подання метаінформації про мультимедійні Big Data, можна стверджувати, що їх семантику можна виразити тільки у вигляді природномовного (ПМ) опису-резюме. Такий текстовий опис семантичного наповнення матеріалу має входити до метаопису Big Data.

RDF (Resource Description Framework) – це інший перспективний підхід до створення семантичних метаданих для різних типів інформації, створений в рамках Semantic Web. Він призначається для стандартизації визначення і використання метаданих ресурсів Web, однак придатний також для опису Big Data. RDF використовує базову модель даних *«об'єкт – атрибут – значення»*. RDF Schema дозволяє визначати конкретний словник для даних RDF і вказувати види об'єктів, до яких можуть застосовуватися ці атрибути, тобто механізм RDF Schema надає базову систему типів для моделей RDF.

Важливою особливістю стандарту RDF є розширюваність: на RDF можна задати структуру опису джерела, використовуючи і розширюючи вбудовані поняття RDF-схем, такі як класи, властивості, типи, колекції. Модель схеми RDF включає спадкування класів і властивостей.

Щоб спростити та уніфікувати створення метаописів ресурсів, користувачам потрібно надати певні шаблони та стандарти опису типових ресурсів. З таких засобів найбільш ґрунтовно розроблено набір елементів для створення метаданих "Dublin Core Metadata Elements" [20].

7 Життєвий цикл аналізу Big Data

Аналіз Big Data відрізняється від традиційного аналізу даних у першу чергу через характеристики оброблюваних даних, таких як обсяг, швидкість і різноманітність. Для задоволення різних вимог щодо виконання аналізу Big Data, необхідна поетапна методологія для організації дій та завдань, пов'язаних із придбанням, обробкою,

аналізом і повторним використанням даних. Традиційний життєвий цикл аналітики Big Data можна розділити на наступні етапи, як показано на рис.1: 1) Оцінювання задачі, яка потребує вирішення за допомогою аналізу Big Data; 2) Ідентифікація даних (внутрішні, зовнішні, адреси місцезнаходження); 3) Збір і фільтрація даних; 4) Витяг даних (отримання, пересилання, запис у сховище); 5) Перевірка й очищення даних; 6) Агрегування й подання даних для аналізу; 7) Аналіз даних; 8) Візуалізація даних; 9) Використання результатів аналізу.

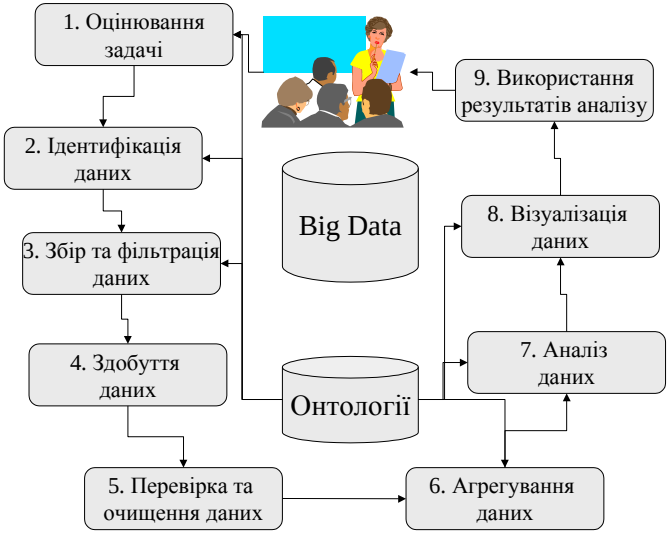


Рис.1 Етапи життєвого циклу аналітики Big Data

Для вирішення поставленої задачі ми змінили деякі етапи життєвого циклу аналітики Big Data, додавши у певні етапи елементи Semantic Web, зокрема, онтологічного моделювання. На етапі *ідентифікація* даних відбувається визначення наборів даних, необхідних для виконання аналітичних проектів (поставлених завдань) і їхніх джерел. Оскільки Big Data на кожному їх етапі супроводжують метадані, які відображають будь-які зміни, які відбуваються з пакетами даних, а також описують характеристики цих даних, то на цьому етапі ми використовуємо семантичний підхід для релевантного вибору Big Data до поставленої задачі. Виявлення

більш широкого спектру джерел даних може збільшити ймовірність виявлення схованих закономірностей і кореляцій у Big Data. Наприклад, щоб дати аналітичний висновок, може бути корисним визначення якнайбільше типів зв'язаних джерел даних, особливо коли неясно, що саме потрібно шукати.

На етапі *збору і фільтрації даних* відбувається завершальне формування пакетів Big Data для цілей поставленої задачі з використанням семантичного аналізу текстових анотацій метаданих і відбору релевантних наборів даних для вирішення поставленої задачі.

Деякі дані, що ідентифіковані як вхідні дані для аналізу, можуть надходити у форматах, несумісних із програмою для роботи з Big Data. Особливо це більш ймовірно для даних із зовнішніх джерел. Етап життєвого циклу *витягу даних*, призначений для витягу непорівняльних даних і перетворення їх у формат, який базова програма для Big Data може використовувати з метою аналізу даних.

Етап *перевірки й очищення* даних, призначений для створення складних правил перевірки й видалення будь-яких відомих неприпустимих даних (дубльовані дані, пропуски даних, надлишкові дані тощо). Для пакетної аналітики перевірка даних і їхнє очищення можуть бути виконані за допомогою автономної операції ETL. Для аналітики у реальному часі потрібно більше складна система внутрішньої пам'яті для перевірки й очищення даних у міру їхнього надходження із джерела.

Етап *агрегування і подання даних* виконує функцію об'єднання наборів даних, які можуть бути розподілені по декількох наборах даних через загальні поля, наприклад через дату або ідентифікатор (ID). У інших випадках ті самі поля даних можуть відображатися в декількох наборах даних. У кожному разі, потрібен метод згортки даних або необхідно визначити набір даних, що представляє правильне значення. Виконання цього етапу може ускладнити через розходження в: *структурі даних* - хоча формати даних можуть бути однаковими, модель структури даних може відрізнятися; семантиці - значення, відзначене по-різному у двох різних наборах даних, може означати одне й те саме, наприклад "surname (прізвище)" і "last name (прізвище)".

Слід відмітити важливість перших двох етапів цього життєвого циклу – постановки задачі, для якої здійснюється аналіз Big Data, та відбір набору Big Data, пертинентних цій задачі. Якщо ці дії виконані невдало, то, незважаючи на складність та ефективність

методів аналізу даних, отримані результати не задовільнять потреби користувача.

8 Онтології та Big Data

В інженерії знань під онтологією розуміється детальний опис деякої проблемної області, що використовується для формального і декларативного визначення її концептуалізації [21]. Часто онтологією називають базу знань спеціального виду, яку можна розділяти, відчужувати і самостійно використовувати в рамках розглянутої ПрО [22]. Те, що онтології є адекватним засобом для опису різних ПрО, є на сьогодні загальновизнаним фактом, а широкий вибір онтологій, доступних через Web, підтверджує популярність цього підходу серед різних груп розроблювачів і користувачів Web-застосувань, в тому числі – і для Big Data.

На сьогодні створена велика кількість різноманітних онтологій, описаних за допомогою різних мов та пов'язаних із найрізноманітнішими ПрО. Ці онтології різняться за багатьма властивостями – обсягом, виразними можливостями, призначенням, ступенем формалізації знань тощо. Саме тому існують різні види класифікації онтологій, які різняться за параметрами, що лежать в основі класифікації. В цілому всі такі класифікації онтологій можна поділити на дві групи – семантичні та прагматичні. Більш детальна класифікація онтологій розглянута в [23]. *Семантичні* класифікації групують онтології за параметрами, пов'язаними зі змістом інформації: ступінь формальності представлених знань; рівень виразності та рівень деталізації інформації [24].

Онтологія ПрО – це та частина знань ПрО, що обмежує значення її термінів, які не залежать від іншої (змінюваної) частини знань цієї ПрО. Таку онтологію ПрО можна розглядати як набір угод про предметну область, а інша частина знань ПрО є множиною емпіричних і інших законів цієї області. Таким чином, онтологія визначає ступінь узгодження значень термінів фахівцями предметної області [25].

У різних джерелах пропонуються різні формальні моделі представлення онтологій Проте всі вони містять множину термінів (понять, концептів), яка може підрозділятися на множину класів і множину екземплярів; множину відношень між поняттями, у якій можуть явно виділятися відношення «клас-підклас», ієрархічні (таксономічні) відношення і відношення синонімії (подоби), а також функції – спеціальний випадок відношень, для яких n-й елемент

відношення однозначно визначається n-1 попередніми елементами; аксіоми і функції інтерпретації понять і відношень.

Для побудови онтологічної моделі Big Data доцільно розділяти множину класів та множину екземплярів класів. Доцільно також розділяти об'єктні відношення – між екземплярами різних класів та відношення даних – відношення між атрибутами екземплярів та їх значеннями. Для опису онтологій Big Data будемо використовувати наступну формальну модель

$$O = \langle X, R, F, T, M \rangle \qquad (1),$$

що складається з наступних елементів:

- $X = X_{cl} \cup X_{ind}$ – множина концептів онтології, де X_{cl} – множина класів, X_{ind} – множина екземплярів класів, таких, що $\forall a \in X_{ind} \exists A \in X_{cl}, a \in A$;
- $R = r_{ier\_cl} \cup \{r_i\} \cup r_{ier\_prop} \cup \{p_j\} \cup p_{ier\_prop}$ – множина відношень між елементами онтології, де
- r_{ier_cl} – ієрархічні відношення між класами онтології – це структури часткового впорядкування з верхнім елементом Thing, що можуть встановлюватися між класами онтології і характеризується такими властивостями, як антисиметричність і транзитивність, $r_{ier\_cl} : X_{cl} \to X_{cl}$;
- $\{r_i\}$ – множина об'єктних властивостей, що встановлюють відношення між екземплярами класів: $r_i(a, a \in X_{ind}) = b, b \in X_{ind}$, $r_i : X_{ind} \to X_{ind}$;
- r_{ier_prop} – ієрархічні відношення між об'єктними властивостями класів онтології ;
- $\{p_j\}$ – множина властивостей даних, що встановлюють відношення між екземплярами класів і значеннями з T: $p_i(a, a \in X_{ind}) = t, t \in T$, $p_i : X_{ind} \to T$;
- p_{ier_prop} – ієрархічні відношення між властивостями даних екземплярів класів онтології ;
- $F = \{F_{cl} \cup F_{prop}\}$ – множина тих характеристик, що можуть використовуватися для логічного виведення над онтологією;
- T – множина типів даних (наприклад, рядок, ціле),

значення з яких можуть приймати властивості даних класів онтології;

- M – множина нелогічних правил ПрО.

Така онтологія Big Data містить клас та виділення типових інформаційних об'єктів з наборами семантичних властивостей (відео, аудіо, потокове відео, частково структуровані дані від датчиків), що відповідають:

- різним форматам пристроїв, що генерують Big Data;
- призначенню цих пристроїв;
- географічному розташуванню;
- часові характеристики;
- достовірність джерела;
- умови отримання доступу;
- обсяги та швидкість оновлення.

Великі дані можуть бути як створеними людиною, так і згенеровані машинами і можуть надходити з різних джерел і представлятися в різних форматах чи типах. Тому онтологія Big Data відображає типові джерела Big Data – від діяльності людей (як окремих осіб, так і організацій) через інформаційно-комунікаційне обладнання (з соціальних мереж, смартфонів, комп'ютерів, касових апаратів, банкоматів тощо) та від автоматизованих програмно-апаратних пристроїв (датчики, сенсорні мережі, відеокамери, GPS, пристрої Internet of Things, автоматизовані виробництва, дрони).

В онтології можуть фіксуватися й параметри якості Big Data – зашумленість, точність, ступінь довіри до джерела, якість сигналу, повнота тощо.

Онтологія дозволяє відображати семантику зв'язків між окремими фрагментами Big Data (часові, просторові, комунікаційні (приміром, інформація від смартфонів, між якими були розмови), за ідентифікаторами пристроїв, за тематикою, призначенням тощо). Нижче наведені приклади елементів онтології Big Data (рис.2), що відповідають різним елементам її онтологічної моделі (1):

- $X_{cl} = \{"Big\_Data\_resources","standard","type","format","metadata\_format",...\}$ - $X_{ind} = \{XXX101,...,MPEG7,...,JPG,...\}$;
- "metadata _ fornat" r_{ier_cl} "format" ;
- $\{r_i\} = \{"has\_type","has\_resource","based\_on",...\}$;
- $\{p_j\} = \{"annotation","size","date",...\}$.

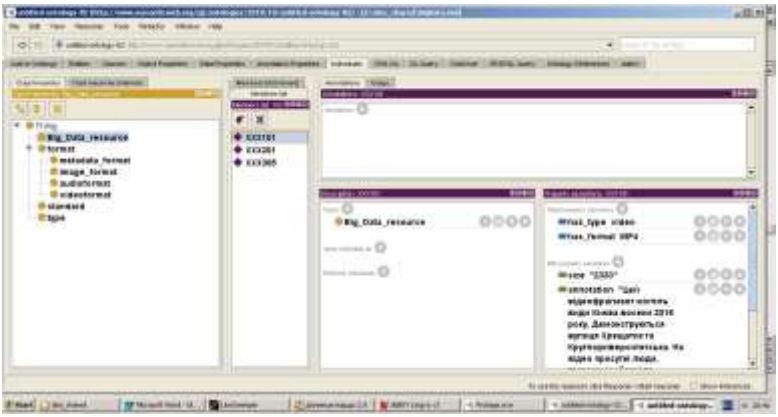


Рис.2 Елементи онтологічної моделі Big Data.

Засоби візуалізації онтології дозволяють легше аналізувати відношення між її елементами (рис.3).

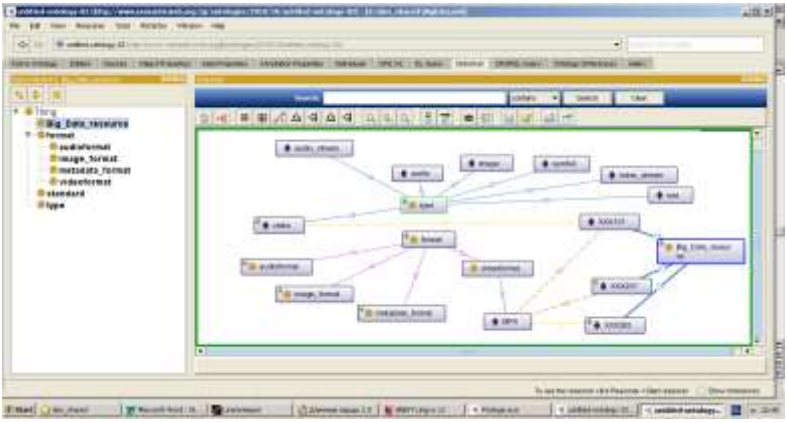


Рис.3 Візуалізація онтології Big Data.

Елементи цієї онтології мають співставлятися з онтологією задачі користувача для пошуку потрібних джерел Big Data.

10 Онтології та Big Data

Щоб використовувати онтологічні знання для співставлення таких ІО, як анотації – неструктуровані природномовні тексти, потрібно забезпечити механізми зв'язування елементів їх контенту з термінами онтології. Пропонується в якості такого механізму використовувати тезауруси задачі, який відображає потреби користувача на основі онтології ПрО, обраної користувачем.

У загальному випадку тезаурус – це словник основних понять мови, що позначаються окремими словами чи словосполученнями, з визначеними семантичними зв'язками між ними [26]. Тезаурус можна розглядати як окремий випадок онтології [27]. *Тезаурус задачі* – це множина термінів ПрО, необхідних для опису та вирішення задачі, для якої користувач намагається за допомогою ССП знайти певну інформацію. Для кожного з них може бути визначена їх вага, що дозволяє охарактеризувати важливість та пертинентність терміну для поточної задачі, та онтологія, з якої імпортовано відповідний термін [28]. Тезауруси дозволяють вирішувати проблеми семантичної розмітки довільних природномовних текстів [29]. Подібність двох ПМ-текстів оцінюється за допомогою функції семантичної близькості між їх тезаурусами.

Для цього виконується порівняння тезаурусу задачі $\mathrm{Th}_{\text{задачі}}$, $\mathrm{Th}_{\text{задачі}} = \left\{ < t_m, w_m >, m = \overline{1,q} \right\}$ та тезаурусів анотацій Big Data з множини I, $I = \left\{ \text{annot}(\text{Big\_Data\_resource}_j) \right\}, j = \overline{1,n}$, підраховується коефіцієнт їхньої близькості $K_j = \sum_{m=1}^{q} f(t_m) * w_m, m = \overline{1,q},$ де $f(t_m) = \begin{cases} 0, & t_m \notin \text{annot}(\text{Big\_Data\_resource}_j) \\ 1, & t_m \in \text{annot}(\text{Big\_Data\_resource}_j) \end{cases}$ та вважається, що $t_m \in \text{annot}(\text{Big\_Data\_resource}_j)$, якщо в анотації ресурсу $\text{annot}(\text{Big\_Data\_resource}_j)$ є фрагмент тексту, який (відповідно до лексичної БЗ) співвідноситься з терміном тезаурусу t_m. Знайдені ресурси впорядковуються в залежності від значень K_j, для подальшого аналізу користувач отримує ті фрагменти Big Data, для

яких значення функції семантичної близькості вище за вказану оцінку.

11 Вирішення омонімії у метаописах Big Data

Через неоднозначність ПМ може виникнути проблема неоднозначної інтерпретації слів, допомогти в вирішення якої має метод *вирішення омонімії.*

Якщо в природному тексті зустрічається слово, що має кілька варіантів смислового значення, то за контекстом треба обрати потрібний варіант (онтологія ПрО) з накопиченням прикладів розпізнавання (множини прецедентів) та алгоритм їх впорядкування (типа традукції) у дерева рішень. Приклади омонімів: гіпербола – „стилістична фігура, в якій перебільшено певну ознаку" і гіпербола – „плоска крива"; кома – „стан непритомності", кома – „пунктуаційний знак" і кома – „газова туманність оболонки ядра комети";

Крім омонімів розпізнавання також стосується різновидів лексичних омонімів: омофонів (слова, що мають однакове звучання, але різне написання: мене – мине, Заєць – заєць, Надія – надія, Селище – селище); омоформ (слова, що мають однаковий звуковий склад тільки в певній граматичній формі: мати (іменник) – мати (дієслово), варта (іменник) – варта (прикметник жіночого роду), шию (від шити) – шию (від шия), три (числівник) – три (дієслово наказового способу), омографів (слова, що при однаковому написанні мають різну вимову, зокрема наголос: замок – замок, дорога – дорога, брати – брати, плакати – плакати).

В мові засобів масової інформації омоніми вживаються в заголовках, у гострих, полемічних, сатиричних, яскраво експресивних текстах. Наприклад, заголовок новин: Голуб і Заєць рятують гусей (депутати Олександр Голуб та Іван Заєць виступили за прийняття закону про заборону весняного полювання).

Джерело інформації для розпізнавання – семантичні (або не семантичні) Wiki-ресурси, а також словники омонімів української (чи іншої відповідної) природної мови.

Контент слова визначається як текст відповідної статті Wiki, з якого здобувають тільки посилання на інші статті.

Процедура розпізнавання омонімів складається з наступних кроків:

1. На вхід схеми поступає текст для розпізнавання.

2. Виконується попередня обробка тексту (препроцесор, блок попередньої обробки).

3. Виконати нормалізацію тексту – перетворення слів в інфінітив, зміна закінчень тощо.

4. Для синонімів у БД уже повинні бути сформовані екземпляри слова-синоніма.

5. Приступити до алгоритму розпізнавання омоніму.

Алгоритм розпізнавання омонімів базується на дереві рішень. Дерево рішень (також називають деревом класифікацій або регресійним деревом) – використовується в інтелектуальному аналізі даних для прогнозних моделей. Структура дерева містить такі елементи: «листя» і «гілки». На ребрах («гілках») дерева рішення записані атрибути, від яких залежить цільова функція, в «листі» записані значення цільової функції, а в інших вузлах – атрибути, за якими розрізняються випадки. Щоб розпізнати (класифікувати) новий випадок, треба спуститися по дереву до листа і видати відповідне значення. Мета розробки алгоритму полягає в тому, щоб створити модель, яка прогнозує значення цільової змінної на основі декількох змінних на вході.

Процес, що йде «згори донизу», індукція дерев рішень, є прикладом поглинаючого «жадібного» алгоритму, і на сьогодні є найбільш поширеною стратегією дерев рішень для даних, але це не єдина можлива стратегія. В Data Mining, дерева рішень можуть бути використані як математичні та обчислювальні методи, щоб допомогти описати, класифікувати і узагальнити набір даних, які можуть бути записані таким чином:

$$(x,Y)=(x_1,x_2,x_3\ldots x_k,Y) \qquad (2)$$

Залежна змінна Y є цільовою змінною, яку необхідно проаналізувати, класифікувати й узагальнити. Вектор x складається з вхідних змінних $\mathbf{x_1,x_2,x_3}$ тощо, які використовуються для виконання цього завдання.

Вирішальне правило (Decision Rule) виду "якщо ..., то...", яке дозволяє ухвалити рішення щодо приналежності об'єкта (слова) до певного класу. Основне застосування вирішальних правил - дерева рішень. У кожному його вузлі міститься вирішальне правило, що розбиває безліч прикладів у ньому на підмножини, асоційовані із класами.

Загальна схема побудови дерева рішень за тестовими прикладами:

- Вибрати черговий атрибут **Q**, поміщаємо його в корінь.

- Для всіх його значень i:
- Залишити з тестових прикладів тільки ті, у яких значення атрибута Q дорівнює i
- Рекурсивно будувати дерево в цьому нащадку

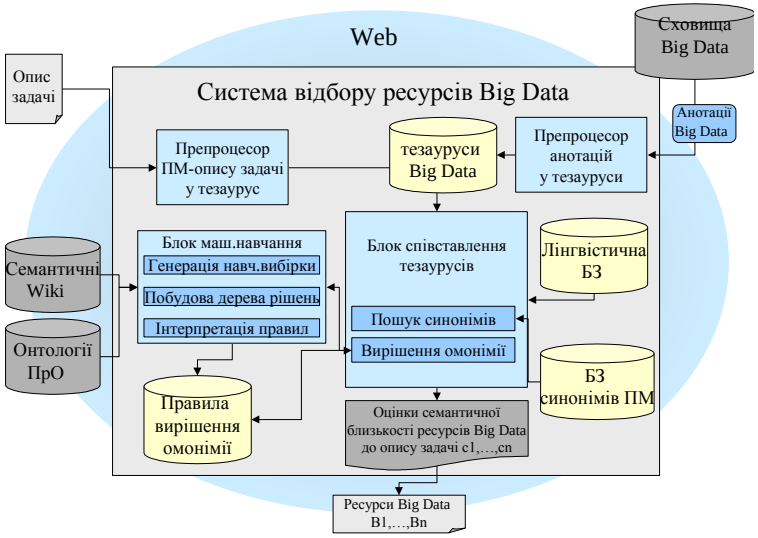


Рис. 4 Архітектура системи співставлення анотацій

Є різні алгоритми індуктивного виведення для вибору чергового атрибуту [12]:

- ID3, де вибір атрибута відбувається на підставі приросту інформації (Gain), або на підставі коефіцієнту Джині.
- C4.5 (поліпшена версія ID3), де вибір атрибута відбувається на підставі нормалізованого приросту інформації (Gain Ratio).
- CART і його модифікації – IndCART, DB-CART.
- Автоматичний детектор взаємодії Хі-квадрат (CHAID). Виконує багаторівневий поділ при розрахунку класифікації дерев;
- MARS: розширює дерева рішень для поліпшення обробки цифрових даних.

6. Розпізнавання омоніму.
7. Машинне навчання (machine learning) для побудови правила розв’язання омонімії.

Необхідно згенерувати навчальну вибірку для слова-омоніма, яке поступає на вхід дерева рішень, – наприклад, слово «коса» має омоніми. Якщо слово «коса» є в тексті із словом «косить», то слово-омонім «коса» є інструментом, Якщо слово «коса» є в тексті із словом «пісок», «море»…., то слово-омонім «коса» є географічним об'єктом. На основі навчальної вибірки генерується вирішальне правило у вигляді дерева рішень «якщо слово $\{c_1,c_2…c_n\}$ поєднується в тексті із словом D, то вибирається слово D". Виконувати ці дії має інтелектуальна система співставлення анотацій (рис.4).

Висновки

Проаналізувавши існуючі засоби анотування Big Data за допомогою метаописів, можна зробити висновки щодо відсутності загальноприйнятого стандарту подання таких метаданих. Тому запропоновані методи аналізу природномовних анотацій є на сьогодні найбільш адекватним засобом співставлення семантики фрагментів великих даних з тими задачами, для рішення яких вони можуть застосовуватися.

Література

1. Ерл Т., Хаттак В., Булер П. Основы Big Data: концепции, алгоритмы и технологии.- Изд.: "Баланс Бизнес Букс", 2017. – 382 с.
2. Марц Н., Уоррен Д. Большие данные. Принципы и практика построения масштабируемых систем обработки данных в реальном времени.- К.: Изд. «Диалектика-Вильямс», 2015.-368с.
3. Bizer C., Boncz P., Brodie M.L, Erling O., The meaningful use of big data: four perspectives – four challenges, SIGMOD Rec. 40 (4) (2012) 56–60.
4. Abbes, H., Gargouri, F.: M2Onto: an approach and a tool to learn OWL ontology from MongoDB database // Madureira, A.M., Abraham, A., Gamboa, D., Novais, P. (eds.) ISDA 2016. AISC, vol. 557, 2017. – Pp. 612–621. doi:10.1007/978-3-319-53480-0_60.
5. Baclawski K., Bennett M., Berg-Cross G., Fritzsche D., Schneider T., Sharma R., Westerninen A. Ontology Summit 2017 communiqué–AI, learning, reasoning and ontologies. Applied Ontology, 2018, P.1-16. – http://www.ccs.neu.edu/home/kenb/ pub/2017/09/public.pdf .

6. Smith K., Seligman L., Rosenthal A., Kurcz Ch., Greer M., Macheret C., Sexton M., Eckstein A. “Big Metadata”: The Need for Principled Metadata Management in Big Data Ecosystems // Proceedings of the Company DanaC@SIGMOD, Snowbird, UT, USA, 2014. – P. 46-55.
7. Dey A., Chinchwadkar G., Fekete A., Ramachandran K. Metadata-as-a-Service //in Proceedings of the 31st IEEE International Conference on Data Engineering Workshops (ICDEW), 2015. – P.6-9.
8. Jeusfeld M.A. Metadata // Encyclopedia of Database Systems, Springer, 2009. – 3. 1723- 1724. – http://www.springerlink.com/content/h241167167r35055/
9. Grotschel M., Lugger J. Scientific Information System and Metadata. Konrad-Zuse-Zentrum fur Informationstechnik, Berlin. – http://www.zib.de/ groetschel/pubnew/paper/groetschelluegger 1999.pdf
10. Halshofer B., Klas W. A Survey of Techniques for Achieving Metadata Interoperability // ACM Computing Surveys, Vol. 42, No. 2, 2010.
11. Metadata Standards and Applications. Introduction: Background, Goals, and Course Outline. ALCTS. – http://www.loc.gov/catworkshop/courses/ metadatastandards/pdf/MSA Instructor Manual.pdf
12. Uniform Resource Identifier (URI): Generic Syntax. – http://tools.ietf.org/html/rfc3986 .
13. Lagose C. Metadata for the Web. Cornell University. CS 431 – March 2, 2005.
14. MPEG-21 Multimedia Framework, Introduction, ISO/IEC, http://mpeg.telecomitalialab.com/standards/mpeg-21/mpeg-21.htm .
15. MPEG-1, ISO/IEC, 1996. – http://mpeg.telecomitalialab.com/standards/mpeg-1/mpeg-1.htm
16. MPEG-2, ISO/IEC, 2000. – http://mpeg.telecomitalialab.com/standards/mpeg-2/mpeg-2.htm
17. Overview of the MPEG-4 Standard, ISO/IEC, 2002. – http://mpeg.telecomitalialab.com/standards/mpeg-4/mpeg-4.htm
18. MPEG-7 Overview, ISO/IEC, 2002. – http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm
19. MPEG-21 Overview v.4, 2002. – http://mpeg.telecomitalialab.com/standards/mpeg-21/mpeg-21.htm.
20. Dublin Core Metadata Elements http://www.faqs.org/rfcs/rfc2413.html .
21. Gruber T., What is an Ontology? – http://www-ksl.stanford.edu/kst/what-is-an-ontology.html.
22. Guarino N. Formal Ontology in Information Systems // Formal Ontology in Information Systems. Proc. of FOIS'98, 1998. – P. 3-15.
23. Никоненко А.А. Обзор баз знаний онтологического типа // «Штучний інтелект» 4’2009. – С.208-219. –

http://dspace.nbuv.gov.ua/bitstream/handle/123456789/8144/27-Nikonenko.pdf.
24. Obrst L., Ceusters W., Mani I., Ray S., Smith B. The evaluation of ontologies // Semantic Web, Springer US, 2007. – P.139-158. – http://philpapers.org/archive/OBRTEO-6.pdf.
25. Гладун А.Я., Рогушина Ю.В. Семантичні технології: принципи та практики. К.:Універсаріум, 2016. – 388 с.
26. ISO 25964-1:2011, Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval / Geneva: International Organization for Standards, 2011.
27. Нариньяни А.С. Кентавр по имени ТЕОН: Тезаурус + Онтология. – http://www.artint.ru/articles/narin/teon.htm.
28. Гладун А.Я., Рогушина Ю.В. Основи методології формування тезаурусів з використанням онтологічного та мереологічного аналізу // Искусственный интеллект, 2008, №5. – С.112-124.
29. Рогушина Ю.В., Гладун А.Я. Використання онтологічного аналізу для семантичної розмітки стандартів з інформаційної безпеки // Информационные технологии и безопасность. Материалы XVII Международной научно-практической конференции ИТБ-2017. – К.: ООО "Инжиниринг", 2017. – С.195-204. – http://its.ipri.kiev.ua/2017itb.pdf.
30. Рогушина Ю. В. Семантичний пошук у Web на основі онтологій: розробка моделей, засобів і методів / Ю. В. Рогушина. – Мелітополь: МДПУ ім. Богдана Хмельницького, 2015. – 291 с.

NETWORK TECHNOLOGY FOR TRANSMISSION OF VISUAL INFORMATION

Bielievtsov Stanislav[1], Igor Ruban[1], Kyrylo Smelyakov[1], Sumtsov Dmytro[1]

[1]Kharkiv National University of Radio Electronics, Nauky Ave. 14, Kharkiv, 61166, Ukraine
stanislav.bielievtsov@nure.ua, ruban_i@ukr.net, kirillsmelyakov@gmail.com, dmytro.sumtsov@nure.ua

At present, the information and communication systems and technologies are developing at a very high rate and are becoming widely available. At this, with every year there are more and more programs appear on the market which allows communication between people at a great distance with the use of network technologies and the Internet. Many already known programs provide intercourse for their users with the use of reliable communication for free or almost free of charge; for example: Skype, Viber, NetMeeting, Net Speakerphone, Team Speak, and Discord. All these programs and technologies are called IP-telephony; it includes a variety of technologies that provide the transfer of multimedia data (voice, video and various multimedia) over computer networks and the Internet. They are based on various real-time streaming protocols. Many companies and users around the world use IP-telephony services to support remote communication and video communication between people in real time. In this regard, the work considers the features of the functioning of various protocols that are used to transmit visual information between network users and users of various programs and subnets. Most of them use the same network protocols, but have different ways of dealing with network problems, such as insufficient carrying capacity or network losses, which in one way or another affect the quality of real time communication, and, as a result, the quality of the displayed visual information. In particular, the features of several most widely used multimedia network real time transfer protocols (video data with sound - video conferencing) are studied under certain limitations and losses. As well, the results of experiments, the obtained estimates of protocols

effectiveness, and the peculiarities in operation of various protocols under the same test conditions are given.

Keywords: *Protocol, Carrying Capacity, Efficiency, Packet Loss, Frame Frequency, visual information*

1 Analysis of efficiency of the key technologies

One of the most important trends in the evolution of modern telecommunications is the development of IP telephony – a multitude of new technologies that provide the transmission of multimedia messages (voice, data, video) over computer networks on the basis of the IP protocol, including local, corporate, global computer networks and the Internet. The concept of IP-telephony includes Internet telephony, allowing organizing telephone communication between subscribers of the Internet, between subscribers of Public Switched Telephone Network (PSTN) through the Internet, as well as telephone communication of subscribers of PSTN and Internet with each other.

IP-telephony has a number of undoubted advantages, ensuring its rapid development and expansion of the computer telephony market. It is beneficial to end users who are provided with a telephone connection at a rather low per-minute payment or for no charge at all. For companies with remote branches, IP telephony technology allows voice communications to be organized using existing corporate IP networks. Instead of several communication networks, one is used. The undoubted advantage of IP telephony over a regular phone is also the possibility of providing additional services through the use of a multimedia computer and various Internet applications. Thus, thanks to IP-telephony, enterprises and individuals can expand communication capabilities by incorporating modern video conferencing, application sharing, etc.

In networks that do not provide guaranteed quality of service, packets may be lost, the order of their arrival may change, data transmitted in packets may be distorted. Under these conditions, various procedures of the transport layer are used to ensure reliable delivery of the transmitted information. When transmitting digital data, the Transmission Control Protocol (TCP) is used for this purpose. This protocol provides reliable data delivery and restores the original order of the packets. If an error is detected in the packet, or the packet is lost, the TCP procedures send a request for retransmission.

For audio and video conferencing applications, packet delays have a much greater effect on signal quality than an individual corruption of

data. Differences in delays can lead to pauses, "lags". For such applications, a different transport-level protocol is needed, which ensures the restoration of the original sequence of packets, their delivery with minimal delay, real-time playback at exactly specified moments, recognition of traffic type, group or two-way communication. This protocol is the Real-Time Transport Protocol (RTP) which regulates the transmission of multimedia data in packets over a computer network at the transport level and is complemented by the Real-Time Control Protocol (RTCP). The RTCP protocol, in turn, provides control of multimedia data delivery, quality of service control, transfer of information about participants in the current communication session, management and identification, and is sometimes considered as part of the RTP protocol [1]. As well, many other protocols are used in IP-telephony.

UDP (User Datagram Protocol) is one of the key elements of the TCP / IP stack. Using UDP, the computer applications can send messages (in this case called datagrams) to other nodes via the IP network without having to send a preliminary message to set up special transmission channels or data paths. UDP uses a simple transmission model, without any implicit "handshakes" to ensure reliability, ordering, or data integrity. Thus, UDP provides an unreliable service, and datagrams may come out of order, be duplicated, or not reach the addressee at all. UDP implies that error checking and correction are either not needed or must be performed in an application. And although time-sensitive applications often use UDP, since it is more preferable to drop some packets than to wait for delayed packets, in real-time systems this is not allowable. If it is necessary to correct errors at the network interface level, the application can use TCP or SCTP designed for this purpose [2].

MPEG-DASH (HLS) (Dynamic Adaptive Streaming over HTTP) is an adaptive streaming technology that provides the ability to deliver streaming multimedia content via an IP-based network by the HTTP protocol. It is the first decision on streaming data transmission with an adaptive bit rate, which received the status of an international standard. The technology involves splitting content into a sequence of short segments, each of which contains a small excerpt of the content. The content itself can be created in several bit-rate options, and alternative segments aligned on the same timeline become available to the DASH client. As the playback goes on, the client selects the next segment to download and play from the available alternatives automatically, according to the network operation conditions. The client selects the segment with the highest bit rate, which can be downloaded and played

back on time, without lagging and buffering. As it plays, the client automatically selects the next segment to download and play from the available alternatives, based on the network conditions. The client selects the segment with the highest bit rate, which is possible to download and play on time, without lagging and buffering.

The specification provides a special format for describing the media stream, MPD (Media Presentation description) which contains information about the segments (timeline, URL, media characteristics such as video resolution and bit rate). Segments can contain any media data, but, in detail, the specification describes two types of containers: an ISO media file (for example, an MP4 file format) and an MPEG-2 Transport Stream.

The technology does not depend on the used audio- or video-codec. As a rule, one or several representations of multimedia files are available (for example, with different resolution or bit rate), and the choice can be made based on the state of the data network, device capabilities or user preferences, thus creating conditions for streaming with adaptive bit rate and best quality. DASH is also independent of the application layer protocols, so the technology can be used on top of any protocol, e.g. such as CCN [3].

The Real-Time Streaming Protocol (RTSP) – is the application protocol which is designed for use in systems which deal with multimedia data and allows you to remotely control the flow of data from the server, providing the ability to execute such commands as start , pause and stop of playback of multimedia content, as well as access to files located on the server by time. It is developed by the IETF in 1998 and described in RFC 2326 [4].

The RTSP protocol does not perform compression, nor does it determine an encapsulation method for multimedia data. Streaming data transmission is not in itself a part of the RTSP protocol [5].

As a transport protocol is absent, by not relying on UDP or RTP for transporting, for providing the content as a stream of one-address data it is possible to use the Real-Time Protocol (RTSP). This is an application level protocol that was created specifically to control the transfer of real-time data, such as audio and video content. It is implemented according to the correction-oriented transport protocol. It supports player control actions such as stopping, pausing, and expedited forwarding in indexed Windows Media files. You can use the RTSP protocol to stream content to computers running Windows Media Player 9 or later or Windows Media Services 9 or later. RTSP is a control protocol that works with the RTP data delivery protocol to provide content to clients.

Windows Media Services implements RTSP via the WMSRTSP server control protocol plug-in. With a standard installation of Windows Media Services, this plug-in is enabled and connected to TCP port 554 [6].

The protocols for streaming data have different implementation principles, strategies for ensuring quality of services and images, as well as methods for transmitting data over the network. It is precisely because of these differences that doubts arise as to which particular protocol is needed and advantageously used for transmitting streaming information.

This study considers a network for testing IP-telephony (video conferencing) with the aim to estimate the data packet losses, which uses various real-time streaming protocols, but under the same conditions. The obtained data are displayed graphically and recorded in the table of characteristics. On the analysis of the obtained data, the practical recommendations as to the use of the considered protocols are proposed, which is typical for studies of this type [7].

2 Design of experiments

It is assumed that the network consists of the following components (Fig. 1):

– Multimedia server, which houses the video file that simulates video conferencing in real time;

– Ethernet Commutator with the function of setting bandwidth and transmission losses;

– a client, which uses a network analyzer to capture traffic and a video player that uses various real-time protocols to receive multimedia information from the server.

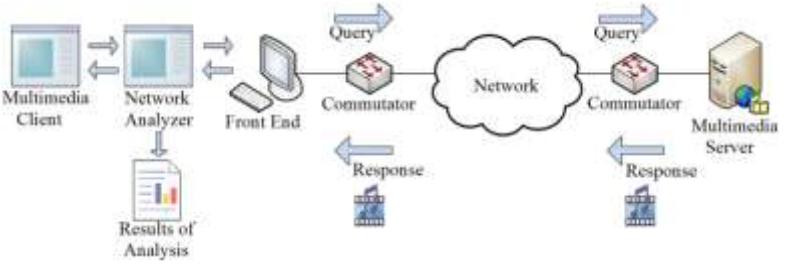


Fig. 1. Simplified presentation of the network [8].

At the multimedia server a video file is disposed which consists of 4 scenes with a resolution of 768×432 pixels, a refresh rate of 25 frames/s, a total stream rate of 1000 Kbps (1 Mbit/s) and duration of 40 s. On the client side, the video file is displayed in the player, while all traffic between the server and the client is captured into the network analyzer buffer. The frame rate in the player was also monitored, which varied depending on the protocol used and various network restrictions. Based on the data obtained, recommendations on the use of the protocols are given.

At first, for registration of changes in frames, a 8% network loss limit was used. Then, a speed limit of 900 Kbps was used. After that, depending on the change in speed from 2 Mbit/s to 800 Kbit/s, the level of losses from 0% to 10% was used; the obtained results are displayed graphically below.

3 The results of the experiment

3.1 Network performance parameters under unchanged conditions

Monitoring of the frame rate during streaming well reflects the playback progress and identifies changes in stream quality that are perceived by a user. In a network in which a packet loss may take place, or which has an insufficient bandwidth, the player can begin to correct the playback speed, and even to reduce it. In such cases, the player can display all the frames included in the video, which increases playback time. The player can also discard image frames; this is perceived by the user as periodic breaks. The breaks may be barely noticeable, or they may be longer periods of freezing frames. They are often caused by a decoder strategy that cancels a damaged video frame and repeats the previous frame until the next valid decoded frame is available. All these player reactions cause changes in the frame rate.

Since the network conditions in the described experiment were simplified, the player basically responded to frame distortions using TCP-based streaming protocols. In most cases, the displayed frames were error free. A user observes the frame errors (reduction in the number of reproduced frames) as periodic breaks or jerks of the image. With a 8%packet loss ratio, the player reproduces video with each streaming protocol tested in the test environment (Fig. 2).

Fig. 2 shows examples of frame rate drops when packet loss is set to 8% and, thus, indicates differences in the interaction of protocols with the FFMPEG player. The first three diagrams (UDP/RTSP, TCP/RTSP, and

HLS) well illustrate the distortions. With RTMP protocol, the playback stops once, and the video freezes for a few seconds. The rest of the sequence is played smoothly. However, in playback with UDP / RTSP, the displayed frames were also strongly distorted [7]. The bandwidth limit, which made visible changes to streaming playback, in the test bench was 900 Kbps for each tested protocol.

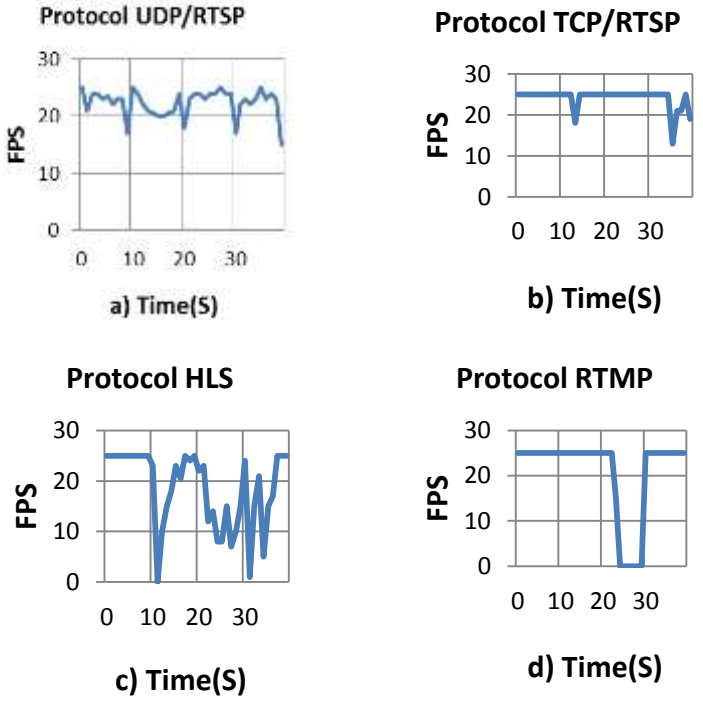


Fig. 2. Falling of frame rate at 8% loss of packets without limiting the rate for protocols UDP/RTSP (a), TCP/RTSP (b), HLS (c), RTMP (d)

Fig. 3 shows examples of frame rates with a 900 Kbps bandwidth. This shows how the RTMP protocol manages to maintain the maximal frame rate. When using the HLS protocol, the playback time increases.

This indicates that the player uses either very short periods of deferral or reduces the playback speed to cope with insufficient bandwidth. In these examples, the RTMP protocol displays most frames: 968 out of 1000. HLS protocol displays 953 frames, RTSP protocol - 873 frames over UDP and 785 frames - over TCP.

Unevenness in the reproduction often occurred near the change of scenes. In addition, the observed quality may vary significantly between scenes. Fig. 4 shows the differences in the scenes during playback. The data were collected from the same test measurements as in Fig. 3. The points associated with the dotted line show the time spent on each scene, compared to the actual duration (40 s), i.e. values greater than 100% indicate that the video was stretched due to a decrease in playback speed or correction of damaged frames due to packet loss.

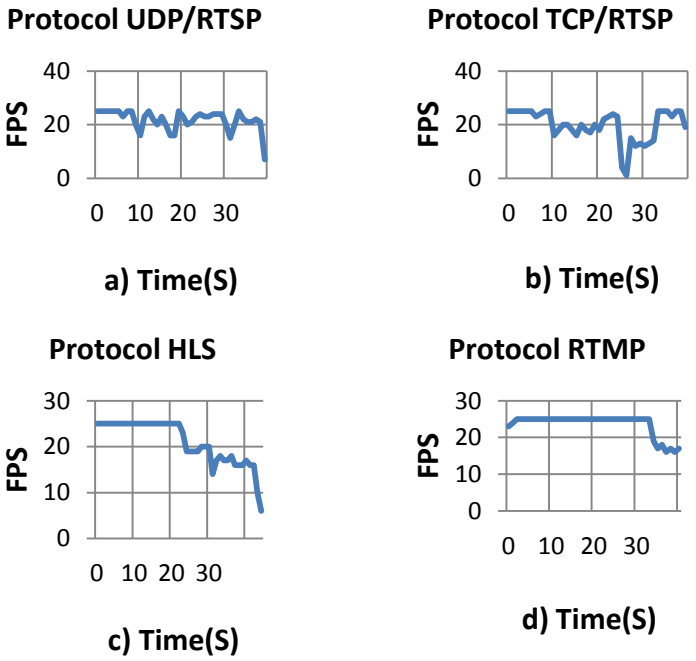


Fig. 3. Frame rate with a rate of 900 Kbps is lossless for protocols: UDP / RTSP (a), TCP / RTSP (b), HLS (c), RTMP (d).

The points associated with the solid line correspond to the displayed frames, i.e. a drop below 100% means that instead of 25 frames/s, fewer frames were displayed due to network losses.

If the streaming conditions were sufficient, both values would remain at 100%. In the RTSP / TCP stream, the third concatenated clip loses

most frames. The streaming gets better in the final scene, allowing you to play almost all the frames, but the increased time spent on it indicates an uneven playback. Listening to the 40-second test sequence took 44 seconds with HLS. With RTMP, the first three scenes were played perfectly [8-10].

From the obtained results, it can be assumed that the number of displayed frames for each protocol directly depends on the introduced network restrictions presented in the following expression

$$F(t) = V \cdot L \cdot P, \quad (1),$$

where $F(t)$ – is the frame rate (fps); V – data transfer rate; L – is the share of lost packets (Packet Loss Rate, omitted in the formula when it is 0%); P – coefficient for the studied protocols (calculated for each protocol): P for UDP/RTSP = 0.00025; P for PTCP/RTSP = 0.00027; P for HLS = 0.00022; P for RTMP = 0.00024.

Exception: at L = 0%. L is omitted and the coefficient K is entered, which has its own value for each protocol: K for UDP/RTSP = 0.1; K for TCP/RTSP = 0.082; K for HLS = 0.107; K for RTMP = 0.11.

In this case the formula takes the form

$$F(t) = V \cdot K \cdot P. \quad (2)$$

Based on the analysis of the results obtained, the Tab. 1 is compiled. According to the data of this table, the final diagram was constructed (Fig. 5).

Table 1. Results of a comparative evaluation of video transmission protocols.

Network restrictions	**Protocol UDP/RTSP**	**Protocol TCP/RTSP**	**Protocol HLS**	**Protocol RTMP**
V = const (1,1 Mbps) L = 8%	22	24	19	21
V = 900 kbps L = const(0%)	22	20	21	24

With a duration of observation interval of 40 s and a frame rate of 25 fps, the total number of frames is 1000. For a confidence probability of $\beta = 0.99$ and the number of experiments n = 1000, the half-width of the confidence interval makes p = 0.005. In other words, with a given number

of experiments, the obtained value of the frequency of reproduced frames deviates from the real by no more than ± 0.005 with a probability of 0.99 [8, 11]. To improve the quality of the results, it is planned to use specialized intelligent systems of frame preprocessing [12-16] and dynamic correction of the transfer process [17].

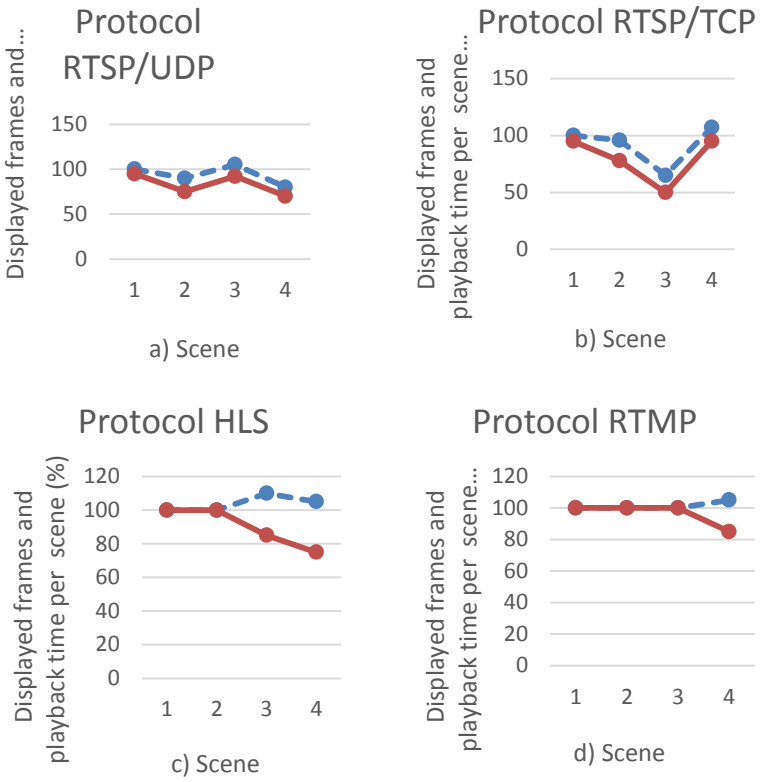


Fig. 4. Differences in scenes during playback at a rate of 900 Kbps without loss for protocols: UDP / RTSP (a), TCP / RTSP (b), HLS (c), RTMP (d)

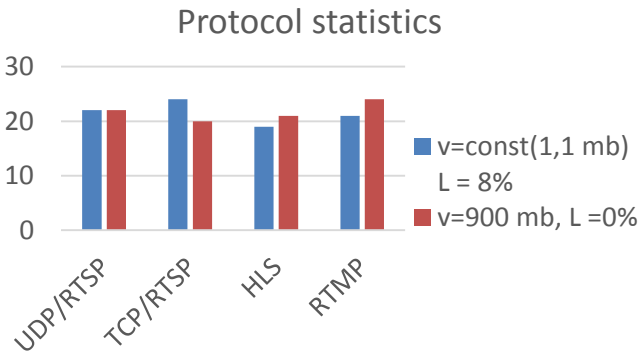


Fig. 5. The ratio of the number of reproduced frames to packet loss.

3.2 Losses at changeable network operation conditions

3.2.1 Losses at a changeable share (%) of packet losses

Fig. 6 shows the number of displayed frames when packet loss was increased from 1 to 10% at a bandwidth of 2 Mbps. In Figs. 6 (a) – (d) the volume of video stream of the player is not limited.

In the RTSP/TCP protocol, the first interference was observed when the loss was 4%. For HLS the limit was 8%, and for RTMP – 11%. For these settings, the RTMP was the only one for which the connection was not interrupted during the streaming, it was able to maintain flawless playback. According to Fig. 6 in the RTSP/UDP stream, the number of displayed image frames and packet losses are directly proportional. However, RTSP / UDP could not cope even with losses of 2% and displayed highly distorted video images. Obviously, this is caused by not providing a support for repeated sending of packets by the UDP protocol.

In Figs. 6 (e) – (h) the size of the video stream of the player is limited. Since the quality of the RTSP / UDP streaming was already poor, changing the queue size did not have a fundamental impact on the quality. HLS protocol cannot display all frames with 3% and RTSP / TCP with 4% ahead.

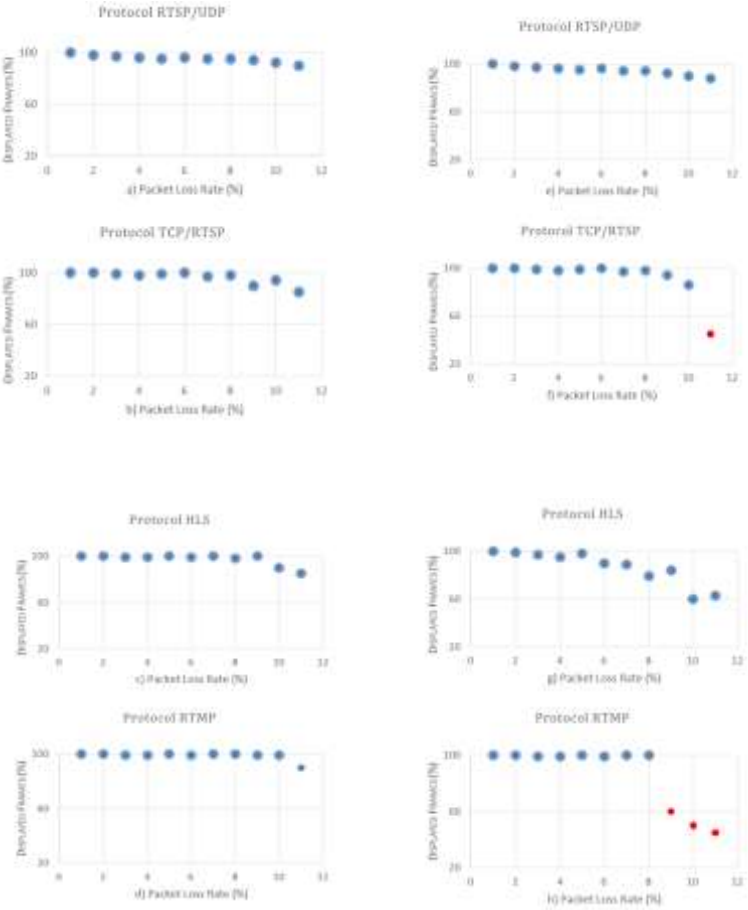


Fig. 6. The ratio of the number of reproduced frames to packet loss. On (a) - (d) the volume of video stream of the player is not restricted, whereas on (e) – (h) – it is limited.

RTMP again supported perfect quality up to 9%. HLS may lose more frames, but it manages to maintain the connection and reproduce the entire sequence with these settings. Since the type of packet lost has a great influence on the propagation of errors and maintaining the quality of streaming, there is also some variation in the displayed image frame metric.

3.2.2 Losses at a changeable rate limit

In order to investigate the response of the player under slow bandwidth conditions, a larger queue depth (of buffer) was used to minimize packet loss caused by the emulator. The bandwidth is limited by the interval of 1.5 to 0.7 Mbps, in descending order. Part of the displayed frames with both queue sizes of the video is shown in Fig. 7. In tests with limited bandwidth, the sequence was reproduced with the same parameters. As in the case of packet loss, frame loss manifested itself as a jerk and short breaks. With RTSP / UDP, a video corruption occurred.

In fig. 7 (a) – (d) video queue volume is unlimited. The streaming was hopeless with all protocols up to 1 Mbit/s bandwidth, which was expected, since the average bit rate of the test sequence was almost the same. Below 1.1 Mbit/s, the HLS protocol began to reduce the number of frames, causing gaps (jerks) of the video. During the experiment, the RTMP protocol lost the connection to the server. The RTSP stream responded by reducing the frame rate more radically, and also lost the connection several times. When the RTP packets were transmitted over UDP, distorted images were also observed at a throughput of 900 Kbps.

In Fig. 7 (e) – (h) the video queue size was limited. With the RTSP / UDP protocol, the player showed highly distorted video images already at a rate of 1.2 Mbps. As the bandwidth decreased to around 1 Mbit / s, the video image was displaying perfectly, with minor jerks. Improving the quality of transients may indicate that at a higher bandwidth, the queue (buffer) of the video player is filled and discards the surplus packets.

When the bandwidth was limited to 800 Kbps, the videos pertaining to the last 10 seconds were distorted. With RTSP / TCPR, resizing the video queue did not significantly affect the performance of the player. In HLS streaming, video twitching (gaps) began with greater bandwidth than with an unlimited queue size (buffer). Similarly, with RTMP streaming, the connection was lost already at 900 Kbps bandwidth [7].

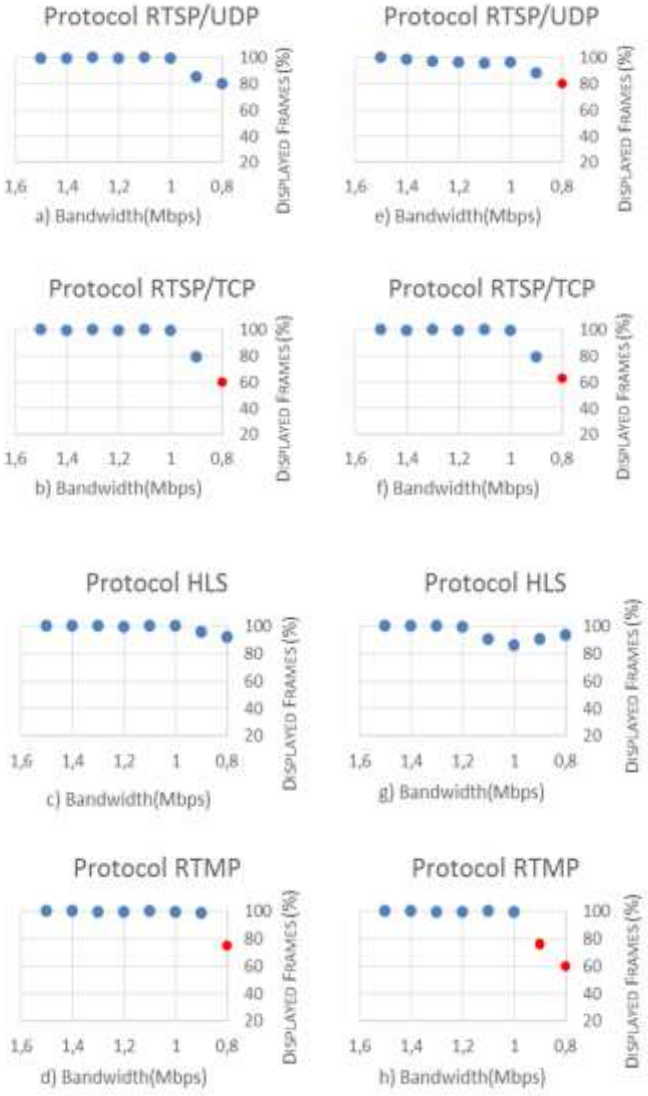


Fig. 7. The change in the proportion of reproduced frames at restricting the rate. On (a) – (d) the volume of the video queue is not limited, whereas on (e) – (h) it was limited.

4 Conclusions

This study compared various real-time data streaming protocols in a computer environment for which packet loss and bandwidth limiting were determined. For these parameters, the boundaries were searched, after which the quality of the streaming begins to deteriorate. According to the simulation results, there are clear differences in the quality of the TCP and UDP streaming protocols. First of all, this is caused by the lack of application layer retransmission facilities in UDP. This means that each lost packet causes gaps (jerks) in the video being played, which are visible to the user, and the number of lost packets can work as an indicator of the quality of the protocol operation. To improve the situation, a re-send mechanism is typically used; for this, for example, a retransmission of lost packets can be used.

In addition to the lack of support for resubmission in the UDP protocol, the player also defines the different sizes of the video data queue in the RTSP as compared to the HLS and RTMP as default values. This was taken into account when modeling. Accordingly, this demonstrates that players compatible with multiple streaming protocols may have different quality indicators.

For protocols based on TCP, the frame rate, the ratio of lost / displayed frames and the monitoring of the playback duration are reflected on the quality of streaming playback, including video slowdown (the appearance of gaps in frames or an increase in the duration of the video due to loss). These indicators can be measured at the application level, and changes in them can be considered noticeable to the user. In particular, changes in frame rates reveal the characteristics of streaming protocols. The ratio of lost / displayed frames shows changes in the frame rate and / or playback duration. Thus, the number of lost / displayed image frames is a metric that generalizes information and is more suitable for assessing the performance of protocols.

In the tests performed, two protocols based on stretching were used. The HLS protocol was the only one that has not lost the connection when network conditions deteriorated. On the other hand, when the size of the video stream (buffer) of the player was reduced, streaming via the HLS protocol was the first to show changes in quality in terms of both loss and bandwidth limitation. This may indicate that this protocol consumes more bandwidth than other TCP-based protocols. The RTMP protocol was retaining the quality somewhat longer than the others. The HLS protocol was the most stable, without breaking the connection during the tests. The RTSP/TCP protocol can cause a greater change in quality in the presence

of packet loss. On the other hand, when the bandwidth was limited, reducing the queue size of the player had no effect on the quality. The RTSP / UDP protocol turned out to be the worst because of the lack of mechanism of repeated sending in the UDP protocol.

References

1. Wikipedia, https://en.wikipedia.org/wiki/RTP_Control_Protocol.
2. Wikipedia, https://en.wikipedia.org/wiki/UDP.
3. Wikipedia, https://en.wikipedia.org/wiki/MPEG-DASH.
4. Wikipedia, https://en.wikipedia.org/wiki/RTSP.
5. Wikipedia, https://en.wikipedia.org/wiki/UDP.
6. MSDN Microsoft, https://msdn.microsoft.com/en-us/library/cc239484.aspx.
7. H.264 QoS and Application Performance with Different Streaming Protocols / Laine, S., &Hakala, I. // 2015 EAI Endorsed Transactions on Future Intelligent Educational Environments, 2015 (3), e3. doi:10.4108/icst.mobimedia.2015.259061
8. Sumtsov D. Development of a method for the experimental estimation of multimedia data flow rate in a computer network / Dmytro Sumtsov, Serhii Osiievskyi, Valentyn Lebediev // Eastern-European Journal of Enterprise Technologies. – 2018. – Vol. 2, N 2 (92). – P. 56-64. – Way of Access: DOI: 10.15587/1729-4061.2018.128045
9. Nbuv, http://nbuv.gov.ua/UJRN/suntz_2017_2_40.
10. S. P. Yevseiev, H. N. Rzayev, S. E. Ostapov, V. I. Nikolaenko Data Exchange Evaluation in global networks based on Integrated Quality Indicator of Service Network // Radio Electronics, Computer Science, Control. – 2017. – Vol. 1. – P. 115-128. doi: 10.15588/1607-3274-2017-1-14
11. H. Venttsel, L. Ovcharov Probability Theory and its Engineering Applications. Moscow: Vysshaya Shkola, 2000. – 480p.

I. Ruban, K. Smelyakov,V Martovytskyi, D. Pribylnov and N. Bolohova Method of neural network recognition of ground-based air objects // IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT), 24-27 May 2018. – P. 589-592. DOI: 10.1109/DESSERT.2018.8409200

12. K. Smelyakov, D. Pribylnov, V. Martovytskyi, A. Chupryna Investigation of network infrastructure control parameters for effective intellectual analysis // IEEE 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering

(TCSET), 20-24 Feb. 2018. – P. 983-986. DOI: 10.1109/TCSET.2018.8336359

13. K. Smelyakov, A. Chupryna, D. Yeremenko, A. Sakhon, V. Polezhai Braille Character Recognition Based on Neural Networks // IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 21-25 August 2018. – P. 509-513.
14. G. Churyumov, V. Tokarev, V. Tkachov and S. Partyka, "Scenario of Interaction of the Mobile Technical Objects in the Process of Transmission of Data Streams in Conditions of Impacting the Powerful Electromagnetic Field", 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 2018. – DOI: 10.1109/DSMP.2018.8478539.

I. V. Ruban, G. I. Churyumov, V. V. Tokarev, V. M. Tkachov, "Provision of Survivability of Reconfigurable Mobile System on Exposure to High-Power Electromagnetic Radiation", Selected Papers of the XVII International Scientific and Practical Conference on Information Technologies and Security (ITS 2017), CEUR Workshop Processing, pp. 105-111, November 30, 2017.

15. S. Mashtalir, O. Mikhnova, M. Stolbovyi Sequence Matching for Content-Based Video Retrieval// IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 21-25 August 2018. – P. 549-553.

ГЕНЕЗА ПРАВОВОГО ЗАБЕЗПЕЧЕННЯ УКРАЇНСЬКОЇ ІКТ-ПОЛІТИКИ

І.Б. Жиляєв, А.І. Семенченко, В.М. Фурашев

Мета статті: на основі аналізу іноземного та вітчизняного досвіду провести ретроспективний аналіз становлення та розвитку правової бази української політики у сфері ІКТ та довести необхідність (важливість) комплексного (інтегрального) правового регулювання ІКТ-сфери, більш широкого, ніж сформульовано у відповідних стратегічних актів з питань нормативного регулювання окремих сфер та напрямів ІКТ-політики.

Гіпотеза дослідження: В Україні сформована певна сукупність ІКТ-політик, кількість яких постійно збільшується, що стимулює активний розвиток їх правового забезпечення.

Вступ.

Основними тенденціями кінця XX – початку XXI століть є:

1) швидкий розвиток ІКТ, що прискорюється;

2) ІКТ-розвиток вимагає створення відповідної системи національного та державного регулювання ІКТ-сфери (артикульованих у вигляді ІКТ-політики);

3) формування сукупності ІКТ-політик, кожна з яких регулює певний сегмент ІКТ-сфери;

4) бурхливий розвиток правового регулювання (нових «масивів» нормативно-правових актів) актуальних суспільних відносин, пов'язаних із впровадженням новітніх ІКТ, що є важливим (вирішальним) елементом кожної ІКТ-політики.

1. Завдання аналізу концепцій українських ІКТ-політик:

Виділити національні / державні ІКТ-політики в Україні – визначивши їх сукупність (ландшафт);

- Визначити відповідні методи та елементи аналізу;
- Дослідити динаміку розвитку правової бази ІКТ-політик за «життєвим циклом» та виявити закономірності їх розвитку.

2. Елементи аналізу української ІКТ-політики

- *Основні:*
  - Концепції (напрями) української ІКТ-політики **K;**
  - Нормативно-правові акти (правова база) **L;**
  - Фактор часу **T;**

- *Додаткові:*
- Завдання (заходи) **P;**
- Ресурси **R;**
- Результати (ефекти) **E.**

Проведення ретроспективного аналізу систем правового забезпечення сукупності українських ІКТ-політик передбачає проведення кількісного та якісного аналізу на основі виділених елементів з формуванням двомірних та тримірних матриць залежності цих елементів.

Українське законодавство визначає «аналіз політики» як методику та практику демократичного урядування, що полягає у всебічному вивченні ситуації та визначенні проблеми у певній сфері (секторі) державного управління, аналізі її причин, визначенні альтернативних шляхів її розв'язання та виборі оптимальних рішень на основі оцінки впливу та із врахуванням позицій різних зацікавлених сторін[1].

3. Аналіз визначення змісту поняття «політика» в українському законодавстві (формулювання)

З розвитком суспільних та економічних відносин у певній сфері настає час юридично формалізувати (легітимізувати) цю сферу, сформувавши єдине бачення на політику. Зазначене зазвичай стимулює розвиток правової бази цієї політики.

В українському законодавстві не існує єдиного стандарту визначення змісту та складових поняття «політика». Кожного разу законодавець, конструюючи поняття «політика» у певній сфері суспільних та економічних відносин, виходить із деяких своїх уявлень щодо її змісту та складових.

Деякі вимоги щодо змісту та технології формування та реалізації державної політики центральними органами виконавчої влади встановлено регламентом уряду, згідно з яким: 1) забезпечення формування та реалізації державної політики у сферах, віднесених до компетенції уряду, здійснюється міністерствами; 2) для вирішення питань суспільно-економічного життя, які потребують визначення концептуальних засад реалізації державної політики, пріоритетів та стратегічних напрямів соціально-економічного розвитку, послідовності дій, вибору оптимальних шляхів і способів

[1] Порядок діяльності груп аналізу політики у центральних органах виконавчої влади, затверджений Наказом Головдержслужби України від 02.04.2010 № 91 http://zakon.rada.gov.ua/laws/show/ z0309-10

розв'язання проблеми, проведення реформ, розробляються політичні пропозиції щодо реалізації державної політики; 3) концепція реалізації державної політики у відповідній сфері базується на оптимальному варіанті розв'язання проблеми та містить розділи: проблема, яка потребує розв'язання; мета і строки реалізації концепції; шляхи і способи розв'язання проблеми; очікувані результати; обсяг фінансових, матеріально-технічних, трудових ресурсів.[2]

4. Концепції (напрями) українських ІКТ-політик К

Поштовхом до формування української ІКТ-політики стало прийняття в 1998 році законів України про інформатизацію. В наступному, у зв'язку із новими проблемами в сфері ІКТ та артикуляцією нових бачень розв'язання цих проблем відокремлювались нові сегменти – формувалась сукупність ІКТ-політик (політика інформатизації була першоджерелом всіх наступних ІКТ-політик). Зазначені політичні рішення щодо концепцій розвиток ІКТ-сфери легітимізувались у базових нормативно-правових актах, які створювали «ядро» відповідної правової бази цих ІКТ-політик.

- **Інформатизація (**Національна програма інформатизації, три закони 1998) – 134 пов'язаних акти;
- **Телекомунікації** (закон 2003) – 1985 актів;
- **Інформаційне суспільство** (Закон 2007; Стратегія 2013) – 76 актів;
- **Електронне урядування** (Концепції 2010; 2017) – 50 актів;
- **Відкритий Уряд** (Ініціатива "Партнерство «Відкритий Уряд» 2012) – 30 актів;
- **ІКТ-безпека: 1) інформаційна** (Доктрина, 2017) – 7 актів; 2) **комп'ютерна безпека** (2001, конвенція), яка у наступному отримала назву – **кібербезпека** (Стратегія, 2016; закон, 2018) – 32 акти; 3) **безпека інформаційних ресурсів**; 4) **безпека**

[2] Регламент Кабінету Міністрів України, затверджений постановою Кабінету Міністрів України від 18.07.2007 № 950 (у редакції постанови Кабінету Міністрів України від 09.11.2011 № 1156) http://zakon.rada.gov.ua/laws/show/950-2007-п

критичної інформаційної інфраструктури держави[3] (захист інформаційно-телекомунікаційних систем об'єктів критичної інфраструктури держави, зокрема – інформаційних технологій та телекомунікацій (електронних комунікацій)[4];

- **Програмна продукція** (Закон, 2012) – 3 акти;
- **Цифрова економіка та суспільство** (концепція, 2018) – не виявлено пов'язаних актів;
- **Пріоритети:** 1) **пріоритетні галузі економіки** (для інвестиційних цілей) – віднесено «виробництво нових та імпортозаміщуючих видів комп'ютерів, електронної та оптичної продукції, машин і устаткування» (закон 2012; постанова 2013); 2) **науково-технічні пріоритети** (закон 2001; постанова 2011) + програми; 3) **інноваційні пріоритети** (закон 2011, державні – постанови 2012, 2017; галузеві – постанова 2017) + державні програми з розвитку ІКТ;
- Галузеві та регіональні (територіальні) політики в сфері ІКТ;
- ...

5. Приклад 1: аналіз розвитку системи правового забезпечення визначеної української ІКТ-політики (на прикладі політики інформатизації)

- Об'єкт: Національна програма інформатизації, 1998
- Термін аналізу: 1990-2018 РР.
- Масив нормативно-правових актів > 3000
- Кількісний аналіз динаміки

«Життєвий цикл» української політики інформатизації

Суспільна увага до певної державної / національної політики призводить до потреби формування відповідної її правової основи. Динаміка формування зазначеної правової бази також (окрім інших елементів політики) демонструє відповідний «життєвий цикл» цієї політики.

---

[3] Див.: Порядок формування переліку інформаційно-телекомунікаційних систем об'єктів критичної інфраструктури держави, затверджений постановою Кабінету Міністрів України від 23 серпня 2016 р. № 563

[4] Примітка: сучасною тенденцією у сегменті забезпечення безпеки ІКТ-сфери є ускладнення та деталізація по окремим (більш-менш автономним) видам ІКТ-безпеки.

Яскравим прикладом є законодавча база української політики інформатизації, яка існує понад 20 років. На рис. 1 представлена нормативно-правова база з питань інформатизації (більш, ніж 3 тис. актів, згрупованих за роком їх прийняття), на якому можна чітко виділити чотири цикли цієї політики: 1990-1995 рр. – становлення; 1996-2003 рр. – стадія росту, 2004-2010 рр. – зрілість, 2011-2018 рр. старіння.

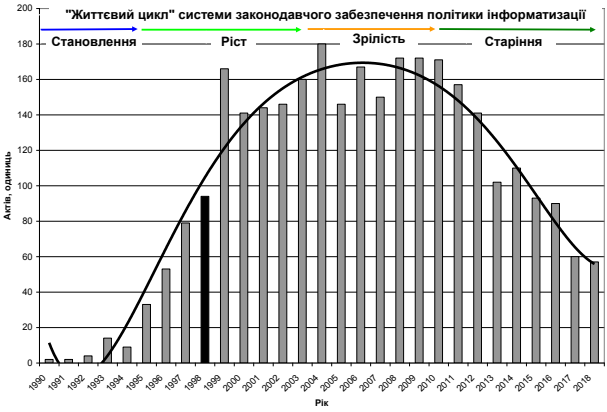


Рис. 1. Чисельність нормативно-правових актів політики що регулюють інформатизацію (з використанням терміну "інформатизація") у 1990-2018 рр. (за роком схвалення). Джерело даних: База даних «Законодавство України» (станом на 18.10.2018)

6. Приклад 2: аналіз розвитку системи правового забезпечення визначеної європейської ІКТ-політики (на прикладі прийняття актів ЄС з інформаційного суспільства чи кібербезпеки)

Наявність «життєвого циклу» ІКТ-політиці притаманно як політикам у конкретних сферах економіки (у нашому прикладі – у сфері ІКТ), так й для конкретних країн.

На рис. 2 розглянуто, як змінювалась правова база європейської ІКТ-політики щодо інформаційного суспільства. В базі даних європейського права EUR-Lex станом на 24.10.2018 зафіксовано 510 документів щодо інформаційного суспільства, з них 403 акти європейського права. Можна бачити, що майже 2/3 документів європейського права щодо регулювання інформаційного суспільства було прийнято протягом 2007-2013 років (на стадії зрілості життєвого циклу цієї ІКТ-політики).

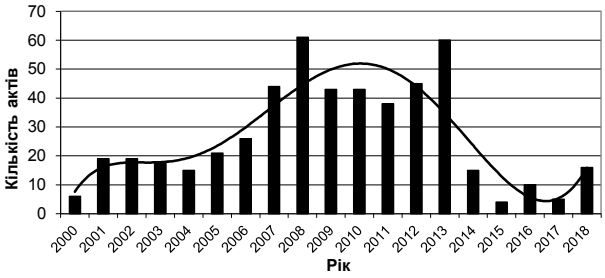


Рис. 2. Кількість актів європейського права з політики інформаційного суспільства (Information society), 2000-2018 рр.
Джерело даних: База даних «EUR-Lex» (станом на 24.10.2018)

По іншому виглядає розвиток правової бази більш «молодої» європейської ІКТ-політики кібербезпеки (Cybersecurity). В базі даних європейського права EUR-Lex станом на 24.10.2018 зафіксовано 668 документів з питань кібербезпеки, з них 5953 акти європейського права. На рис. 3 можна бачити, що майже 2/3 документів європейського права з кібербезпеки було прийнято протягом 2017-2018 років (на стадії бурхливого росту життєвого циклу цієї ІКТ-політики). При цьому актуальність цієї проблеми примушує законотворця все інтенсивніше врегульовувати суспільні відносини у цій сфері.

Висновки

1. Україна достатньо динамічно формує ІКТ-політичний ландшафт, його правову базу.

2. Однак, сучасний ландшафт сукупності українських ІКТ-політик є:

- фрагментарним – політики начастую дублюють одне одну, використовують відмінні терміносистеми (несистемність та адитивність);
- таким, що має «розриви» між ІКТ-політиками та «розриви» у ланцюжку їх життєвого циклу (від старту – створення до фінішу – утилізації);
- таким, що динамічно змінюється, «відторгаючи» попередні» ІКТ-політики, начастую – невиправдано;
- змінним під впливом нових іноземних концепцій та ініціатив (залежним від ресурсів донорів);

- проблемними з точки зору результативності / ефективності: існує «невідповідність амбіцій амуніції» (ресурсів, що виділяються поставленим цілям політики, втрата довіри до концепцій політики.

3. Бракує досліджень природи нових правових масивів, пов'язаних з появою нових ІКТ-політик; механізмів формування таких масивів, їх впливу на суспільні відносини, право, його систему та структуру, правозастосовну практику.

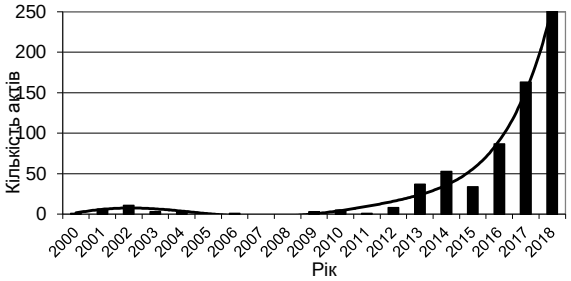


Рис. 3. Кількість актів європейського права з політики кібербезпеки (Cybersecurity), 2000-2018 рр. Джерело даних: База даних «EUR-Lex» (станом на 24.10.2018)

Пропозиції

- розглянути необхідність формування інтегрованої національної ІКТ-політики; уніфікувати (формалізувати підходи) до змістовної частини формування та реалізації окремих ІКТ-політик;
- систематизувати нормативно-правові актіи щодо регулювання ІКТ-сфери (усунення дублювання та суперечностей, формування єдиної терміносистеми, відміна застарілих актів тощо);
- забезпечити випереджаючу розробку правового регулювання суспільних відносин, пов'язаних із появою новітніх та прогнозованих ІКТ[5], активізуючи імплементацію норм європейського права.

[5] Традиційно теорія права відносить до сфери правового регулювання суспільні відносити, які: 1) можуть і повинні бути є врегульованими; 2) можуть і повинні бути та є неврегульованими. Цифрова епоха потребує правового врегулювання нової групи відносин – які повинні бути, але на сучасному етапі не можуть бути врегульовані правом.

SEMANTIC OPTIMIZATION OF WEBSITE CONTENT BASED ON USER PREFERENCES

Budko Alexandr[1], Igor Ruban[1], Kyrylo Smelyakov[1], Maslovsky Vladislav[1]

[1] ***Kharkiv National University of Radio Electronics, Nauky Ave. 14, Kharkiv, 61166, Ukraine***

oleksandr.budko@nure.ua, ruban_i@ukr.net, kirillsmelyakov@gmail.com, akmyto@gmail.com

This work deals with the actual problem of semantic optimization of the content of websites in accordance with the individual preferences of the user and with respect to the basic concepts of conversion and site metrics. The main mistakes of web developers are considered which pertain to developing of web site layouts that affect the conversion significantly, and reduce noticeably the user activity and the percentage of their involvement in business processes, services provided, product sales, which were assumed by website owners for operations of clients with their products. The effectiveness of the use of key tools for analyzing the characteristics of web applications has been studied; in particular – on the basis of analysis and generalization of the results of numerous experiments related to the manifestation of user activities in various fields of activity in cyberspace. The general problem of compiling algorithms for complex semantic optimization of web site content in accordance with the existing standards and individual preferences of user has been stated and decomposed. The basic behavioral models are described, as well as the general patterns and trends in users' behavior depending on their age, gender, web pages content allocation, amount of information on a page, presence of graphic elements and many other factors. On the ground of these studies an algorithm for semantic optimization of web applications based on user preferences has been developed. The proposed algorithm has been analyzed with respect to the effectiveness of its application to finding the problem areas of a web application. It is also described the scheme of actions which one should follow for an effective application of the proposed algorithm. As well, a system of practical recommendations has been proposed which allow a web developer to create applications with a high level of conversion and a low number of failures.

Keywords: *semantic optimization, website, conversion, click heat map, site map.*

1 Analysis of the current state of the issue

Modern Internet applications contain constantly updated and diverse content. It often happens that a user, opening a website for the first time by using a link in a search service, stumbles upon a not entirely user-friendly interface. On this site there is information he needs, but it is usually at the bottom of the page, and for reaching it you need to scroll for a long time, whereas the information itself is replete with redundant links. Flickering promotional offers, next to the text, constantly distract from reading interesting information. To view the comments, you need to pass registration which is framed in a terribly inconvenient form with a poorly readable CAPTCHA-like test. Most often it happens that even at a small hindrance with data entry the user closes the tap already after 5 seconds and mentally places the site on the list of "unfriendly" ones. Based on this, it can be concluded that the number of visits to such a site is greatly reduced due to the fact that useful information is presented to the user in an unstructured and poorly formatted form [1]. These sites include many blogs, social networks, online auctions, photo and video hosting. In such a situation, it becomes difficult for users to navigate the variety of information presented both on one web page and on the whole website. Imagine that the user's behavior on the site is similar to the behavior of a laboratory mouse in a maze: the mouse must go through the maze and find a piece of cheese. From the point of view of the creator of the site, the task is to create such a labyrinth being convenient for a mouse, which allows it to find its piece of cheese without fail. So, as the problem embraces three parts: site search, site usage and site content [2], it has to be considered in a complex of its aspects.

The main tasks of the work are to summarize and analyze the results of experimental research in the field of semantic optimization of websites content in order to develop an appropriate algorithm and formulate practical recommendations to help a web developer in creating application layouts and filling it with content so as to best fit the preferences of the target audience and Increase website conversion rate.

The conversion is understood as the ratio of the number of visitors, who has performed some goal action, to the total number of visitors. Most often, conversion is measured in percents (%); for this, the ratio is

multiplied by 100%. So, the formula for calculating the site conversion is as follows

$$\text{Conversion} = [\text{N of Goal}] / [\text{N of Visitors}]^* 100\%. \quad (1)$$

Here [N of Goal] is the number of visitors who have achieved the goal, and [N of Visitors] is the total number of site visitors [3].

It is proposed to search for the solution of the problem with the help of the algorithm proposed in the work, the idea of which is to apply the following basic steps.

1. Choice of tools for collecting statistics on user behavior and evaluating the metrics.
2. Search for places, where visitors actively demonstrate the pattern of their behavior under normal conditions, for obtaining the most relevant data (social networks, blogs and some other sources).
3. Analysis of the accumulated experimental data by generating usage reports and finding patterns of user behavior in various web layouts; generalization of the results in tabular form; finding ways to correct the basic errors of web development.
4. Development of a web application semantic optimization algorithm which is based on user preferences.
5. Formulation of recommendations for increasing the conversion of web applications and improving the quality of content in accordance with the user's preference.

2 The effectiveness of modern solutions and tools for semantic optimization

Today, there are a number of ready-made solutions and tools that can help in collecting statistics on user actions and analyzing the main difficulties that he may encounter. In this study, we deal with such tools as Journey map, Click heat map, and Usability testing [4].

2.1 Journey map (conversion map)

Consider firstly the journey map. A customizable journey map in general allows us to see the routes that the users use on the site, to assess the correctness of the site structure and the adequacy of internal links (Fig. 1). In this, we seek for the answers to the questions: why the users

are not being registered; what are the most popular routes of passage through the site, whether they are what were meant.

2.2 Click heat map

Further on, it is required to reveal where the user clicks, after he got to the page he needs. Visualizing the clicks "as if sitting behind the user's shoulder" is another way to see the users' behavior on the site (Fig. 2). The click heat map helps to understand this. Firstly, it answers the question – have you successfully arranged the most important parts of the site material, or not. Also, how your linking works: whether the links are being clicked, or forgotten?

Fig. 1. Google Analytics Journey Map [4].

Fig. 2. Click heat map of Google Analytics [5].

2.3 Eye-tracking testing

The third tool is the Usability testing. This technology is very similar to the click map. The heat map, which is obtained as a result of usability testing, is different in that it is built not on the basis of mouse clicks, but on the basis of the movement of the user's eyes, which are recorded by a special camera (Fig. 3). The so-called eye-tracking testing allows you to get even more data as to a successfulness of placement of material on the site; at this, a standard heat map, built in analytics based on mouse clicks, is often called a cheaper alternative to eye-tracking.

2.4 Choice of semantic optimization tools

One of the main advantages of maps for a mouse in relation to a map for eyes is that when you track a mouse, you get data from actual visitors. Meanwhile, in the case of tracking the eyes, you use a special group of people who are often extracted from a normal environment, so they can generate distorted results. On the other hand, the use of eye-tracking may give you the results with an accuracy up to 100% relative to the probationer, while the accuracy of mouse analyzer makes about 85-90% [7]. In order to make the final choice, we have conducted an experiment in which we studied each of the specified tools, determined the number of metrics pertaining to the subject, and analyzed the pros and cons of each. Accordingly, the summary characteristics for various semantic optimization tools are presented in Tab. 1. We made our choice in favor of heat maps, since it is important for us that the results obtained should correspond maximally to the behavior of the average user.

Fig. 3. A Heat map of Google Analytics eye-tracking [6].

Based on a comparative analysis of various tools, it is advisable to opt for two tools – Google Analytics and Yandex Metrics. Google Analytics has one drawback – the absence of all information about the key queries from the search traffic that comes from Yandex. This is due to encrypting of HTTP referrer for all analytics services, except Metrics. Therefore, if in Yandex Metrics, you will receive data on keywords for 2/3 of search queries, then in Google Analytics – for 1/3. Therefore, it is quite expedient to use two services simultaneously, in parallel to each other, in order to obtain more accurate data. Consider now the key results for the click heat maps of the selected tools.

Table 1. Comparison of services for visitors' behavior analysis.

Name	N. of metrics	Difficulty of use	Price	The need to download
Google Analytics	12	Above the average	For free	No
Yandex.metrics	10	Above the average	For free	No
Clicky.com	8	Low	For free, $9.99 a month	No
Piwik.com	4	Low	For free, 65$ a month	Yes
Woopra.com	7	Average	78$ a month, Demo version	Yes, Mobile app
Mixpanel.com	6	Average	For free, Partnership 150$ a month	Yes, Mobile app
KISSMetrics.com	7	Average	200$ a month, Demo version	Yes, Mobile app
Gosquared.com	6	Low	18$ a month, Demo version	No
Chartbeat.com	5	Average	9.95$ a month, Demo version	Yes, Mobile app
Goingup.com	6	Average	For free	No
Openwebanalytics.com	6	Average	For free	Yes
RJMetrics.com	7	Above the average	500$ a month	No

3 Obtaining and analyzing the results of experiments based on click heat maps

3.1 View page content by segment

As a result of the Nielsen Norman Group experiment (hereinafter NNG) [8], it was found that users spent about 57% of the time browsing a page over the bar. The content above the bar receives to date the largest proportion of the viewing time. About 74% of the time was spent on the first two areas of content (information above the bar plus the window immediately below it).

The remaining 26% was spent in small steps further along the length of the page. The results of the experiment are listed in Tab. 2. It is clear that not every page has the same length. To determine how people divide their attention along the page (regardless of its length), the NNG divided the pages into 20% -segments making one fifth of each page.

According to the Tab. 3, where the data of this experiment are presented for the general websites, we see that more than 42% of the viewing time was devoted to the top 20% of the page, and more than 65% - to the top 40% of the page. On the search results pages that the team highlighted in the 2010 results, 47% of the viewing time was devoted to the top 20% of the page (and more than 75% - to the top 40%). Apparently, these data reflect the desire of users to look only at the top results.

Table 2. Duration of viewing the page content (in separate areas).

Page segment	Percent
First	57%
Second	17%
Third-Ten	7%
Fourth	5%
Fifth	3%
Sixth-Ten	7%
Last	4%

Table 3. Duration of viewing the page content (in segments).

Page segment	Percent (search system)	Percent (Web site)
20%	42%	47%
40%	24%	27%
60%	16%	11%
80%	11%	8%
100%	8%	6%

If people only look at the content above the bar – within the first area – the information at the top of the screen gets more attention than the information at the bottom. As you can see from the results given in Tab. 4, more than 65% of the viewing time above the bar was concentrated in the upper half of the viewing window. On the search results pages, the top half of the first screen received over 75% of the viewing time over the bar.

Table 4. Duration of viewing the page content in the first area.

Content segment	Percent (search system)	Percent (Website)
20%	42%	47%
40%	24%	27%
60%	16%	11%
80%	11%	8%
100%	8%	6%

3.2 Blocks that look like ads

In another experiment, the NNG [9], by using a heat map, fixed the user's attention to blocks that might look as advertisements. As a result, it was noticed that some participants learned to skip the ad presented at the top of Google search results, although its visual design is far from the traditional advertising banner. Ignoring ads is a scientific behavior that, like many other user behaviors on the Internet; classic examples include searching for a company logo in the upper left corner of the page or searching for global navigation at the top of the page.

This viewing area shows that on the Google search engine results page (SERP), the user was not looking at the entire first "result," the advertisement (Fig. 4). Today's ads can appear anywhere on a web page, and users are aware of this fact. Thus, they try not to waste time on advertising, even when advertising is displayed in content areas. In the workspace, the embedded ads are relatively easy to ignore because they are very different from the surrounding page elements. For example, 26 participants of the experiment (Fig. 5) who were trying to learn about the dance education of Mikhail Baryshnikov, ignored the sentences that appeared in the text.

Fig. 4. Heat map of Google Analytics views [9]. Here, each red dot is a fixation of the position of the gaze of one user; the lines connecting them are fast eye movements or moments during which the user sees nothing.

The main reason they ignored these blocks was because the sentences looked different than the text and the photo on the site. In particular, the ad stood out because of several qualities: a small rectangular figure in the middle of the text, unusual formatting, colored (blue) background on a white page, text embedded in the image.

Each of these features warned users that the rectangle is an "advertisement", so they can safely ignore it.

Fig. 5. Heat map of clicks Google Analytics [9].

Content placed on the same part of the screen as the advertisement is often considered an ad, and is also ignored. This simple consequence of the well-known principles is the law of closeness of the Gestalt: elements close to each other are considered part of a group and, therefore, are associated with a function. Because people analyze various items within the screen area, they form a mental model of the content available there, based on the informational smell of the items they visit. Thus, if one of the elements seems completely irrelevant, they often assume that the whole section is not related to their purpose and stop exploring other subjects.

The participant in the experiment, who was studying how to remove a spot, immediately looked at the right-hand block, presumably deciding that it contains only ads. They actually contained sponsored stories, but it also contained useful, fun videos that show how to make various handicrafts, such as crocheting a rug, or creating a magnetic frame. The user was deceived by an ad located in the same section of the page and did not view other content in this block.

3.3 Posting full articles and their summaries

The following experiment focuses on [10] how people read the "official" blogs of companies, government agencies and large non-profit organizations. As a result, the experiment showed that users are not inclined to read the full articles on the first page, but prefer to view their short description. The following heat maps show how users read the first page of several corporate blogs (Fig. 6).

The first two examples show blogs with full articles on the first page. In both cases, users viewed the first article, but did not look further. If your first article has not interested the users, you would likely lose them since they had wasted the interest during browsing the first paragraph. This unproductive interaction with the web page has definitely exhausted the user's interest in the site.

After spending so much time viewing one article, he said he was disappointed that the site did not offer a summary for other articles. Since the user did not want to spend any more time studying full articles, he left this weblog after the first visit. On the second heat map, the user carefully read about 4 paragraphs, and then looked up to 3800 pixels at the top of the page. The following two examples of blogs contain a summary on the main page.

On the first blog, the user viewed 10 resumes that were shown on this relatively short page. In the second, the user viewed all 5 resumes on this even shorter page.

The last example shows a hybrid approach: many publications are short and shown in full, while more detailed articles are presented in a short description with reference to the full text. Here the user has viewed 11 messages.

Fig. 6. The click heat maps of Google Analytics [10].

The last blog also shows how photo blogs can draw users down the page: it is easier to scan a long series of photos than read. In a similar

blog, the user has studied more than 12,000 pixels by allotting 3-4 viewing for each of the many photos, while almost not reading the text.

Based on the results of this experiment, it can be seen that in corporate blogs, resumes usually dominate in relation to full articles, since they provide users with a wide range of topics. Offering more topics increases the likelihood that the user will continue to search and/or find something that really interests him, and therefore he will click to learn more.

In full-text articles, the initial topic may not be of interest to many users, and few will scroll down to see subsequent topics that would be of interest to purchase the products offered. It rarely happens that every thing a blog tries to demonstrate will be of great interest to all customers. Probably, a too wide range of products and an abundance of topics with offers will only interest individual users.

3.4 A brief overview of Additional Experiments

Inspired by Dan Ariely's book "Predictably Irrational," Robert Stevens did a test with 60 random people to see how relativity affects daily decision making. As a result of this test, it was discovered that people do not make decisions based on complete information about the world. We make decisions based on the information we have at the time of their implementation. People have been tested with two different cocktail menus. In the first case, only the prices of cocktails with discount were visible. In the second case, indicated the price and discount. And although the price of discount cocktails remained the same, people were more satisfied with their purchase when they also knew about the initial price – 2.4 versus 1.7 on the scale of identity.

To find out how users behave after testing a landing page using heat maps, it was studied the extent to which color contrast affects visitor behavior. It was found that pricing information on the main page attracted the most attention because of its color contrast with the surrounding space, diverting attention from the rest of the page. After a small designer makeup of the main page, the analysis showed that the site scanning diagram began to correspond to what the company needs [11].

When testing 257 correspondents in a remote user test, the failure rate for tasks was 1.9 times higher for people older than 55 than those who were under 25 years old. Almost twice as many older people failed or abandoned this task, compared with young people. Older people are also slower on the Internet. Compared to the youngest participants, the oldest took 40% more time to complete the task.

In one of experiments, the subject spent 10% more time, looking only at photos of the staff, than reading the content that was on most web pages. The study also showed that the left side of the website received 69% of the viewing time. People spent more than twice as much time looking at the left side of the page.

4 Web-application semantic optimization algorithm

For the most effective solution of the problem set in the work, it is advisable to combine individual optimization elements into one algorithm for semantic optimization of a web application. For this, on the basis of the conducted research, the following algorithm is proposed.

The first step is to identify the problem area of the application with a low conversion rate (1), for example, a site that has a small number of registered users.

The second step is the choice of a semantic optimization tool: by the specified causes we stopped on Google Analytics heat maps and Yandex Metrics.

The third step is to collect statistical and visual data. In order to find the "bottleneck" of our application, it is necessary to evaluate various metrics for using the website: depth of viewing, visit duration, source of journey, device type, user's geographical position, failure rate. Thus, Tab. 5 and Tab. 6 present the conversion statistics we have collected.

Table 5. The total time of visiting the site in various browsers.

Browser	Sessions
Chrome	35.87%
YaBrowser	16.06%
Firefox	14.84.%
Opera	14.34%
Internet Explorer	9.92%
Safari	4.97%

The first of them shows the time users spent on our website using a specific type of browser. The second table gives the statistics of using the web application on various types of mobile devices. Based on this data,

we can determine the browser and mobile device, where, most likely, there may occur a vulnerability for further testing.

Table 6. Statistics of use of the web application on mobile devices.

Mobile Device	Sessions	Average duration, minutes	Failure rate,%	Registration
Apple IPad	24.44%	01:30	67.52%	2.64%
Apple IPhone	22.50%	02:22	75.71%	1.16%
Samsung	35.5%	04:61	79.22%	5.19%
Google Nexus 7	0.59%	00:22	79.43%	2.84%
Opera Mini for S60	0.57%	00:39	73.57%	4.41%
Not installed	16.48%	00:58	76.10%	3.38%

The fourth step is to analyze the accumulated data. As you can see, the most popular browser is Chrome, and the least amount of time users spend in Safari. The reason for this may be both the low popularity of the browser and the poor browser support of our markup, so it is necessary to revise the layout of our page. Perhaps some of the content is not displayed correctly or does not scale well within this type of browser. Next, analysis of the following table, derived from Google Analytics statistics, shows that the smallest number of registrations falls on the Apple IPhone, although the mean duration of session for this device is close to general average value. The journey map shows that many people come to the registration page, but not all are registering; the problem is found in the form itself: it may be inconvenient, too long, or simply not sent to the server.

Next helpful instrument is the heat map of clicks. With it, we have found that the long registration form dumped the fields entered into it after an incorrectly entered Captcha-like check data. The fifth step is to change the vulnerable area based on user's preferences. After we have removed the Captcha-like check and shortened the form by two times, the number of registrations increased by one third (Fig. 7).

As the result, by grounding on the results of the conducted research, we propose the following key recommendations which, as the results of the experiments show, are advisable to be followed in formation of webpage layouts and their content.

1. The most important information for the target audience of the website should be located in the upper part of the page.

2. A visitor skips through the middle part of the page, but his attention increases again in its lower part. Inserting appealing and informative content at the end of a web page can increase user interest.

3. Use extensively the visual presentation of information in order to emphasize the most important. Use this principle so that visitors can quickly find what they are looking for, or what you want to sell the most.

Fig. 7. Heat map of registration form after our corrections [12].

4. Use photos for capturing the attention and guiding. Use photos of real people: humans respond well to images of real people.

5. When an element of your website looks like an advertising banner, it will receive the least attention. A notice should not appear at the top of the page or in the right block, or otherwise it might be ignored with a high probability.

6. For Blogs a resume is better than a full article.

7. People read your content in an F-pattern mode. Therefore, in the first two textual paragraphs of the website, you must specify the most important information. Use subheadings, markers, and paragraphs to structure content and make it readable.

8. People spend more time looking at the left side of your page. Use this side of your website to display the most important information, e.g. the discounts on services. Place here the logo of your company.

9. Displaying a reduced price next to the original one will increase the satisfaction from the purchase.

10. Use color highlighting of webpage elements to direct visitors to your preferable business proposals or products you are selling.

For efficient implementation eye-tracking technology and to improve the quality of data analysis, it is planned to use specialized intelligent systems of frame processing [13-17].

5 Conclusions

In this study we have analyzed the up-to-date trends in web application development, investigated the effectiveness of basic tools for semantic optimization of web applications, collected, analyzed and generalized the experimental data from various studies of heat maps; on this basis some generalized models are compiled and the behavioral preferences of modern users are formulated are formulated, an algorithm for semantic optimization of low conversion web application is developed, and a series of practical recommendations to website developers are proposed.

The novelty of this study is the proposed algorithm for semantic optimization of web applications. Through in-depth analysis of data of diverse and numerous experiments on the study of user behavior on the website, general concepts and rules were derived that are worth to be followed when building a web application layout; as well, the main stages of semantic optimization of web vulnerabilities were formulated.

The proposed algorithm allows us to collect structured data using a broad range of available tools, by extracting statistical data on the use of the site in order to detect problem areas or to increase the conversion rate. Using this data, a web developer can easily form a certain behavior model of an ordinary user of his application; see how he works with it, what he most often uses or what he does not notice at all. Thus, having collected such information about its visitors, a web developer can quickly see the vulnerabilities of his product and modify it in accordance with the preferences of users within the aim to increase the number of visits or the profits from sales on the site. The prospect of research is the creation of a special product that can automatically analyze user behavior during a session, learn and modify content based on his preferences.

References

1. Andrew B. King Website Optimization. – O'Reilly Media, Inc., 2008. – 367p.

2. Enge E., Spencer S., Stricchiola J., Fishkin R. The art of SEO (2nd ed.) – Sebastopol, CA: O'Reily, 2012. – 714p.
3. Agichtein E., Brill E., Dumais S. Improving Web Search Ranking by Incorporating User Behavior Information // SIGIR 06 The 29th Annual International SIGIR Conference Seattle. – 2006. – P. 19–26.
4. Habr, https:// habr.com/company/altweb/blog/224847.
5. Windata, http://windata.ru/itnews/net/analiz-i-testirovanie-yuzabiliti.
6. Flywebmedia, http:// flywebmedia.com/eye-tracking.
7. Porter J. Designing for the social web. – Berkeley, CA: New Riders, 2008. – 201 p.
8. Nngroup, https://www.nngroup.com/articles/scrolling-and-attention.
9. Nngroup, https://www.nngroup.com/articles/banner-blindness-old-and-new-findings.
10. Nngroup, https://www.nngroup.com/articles/corporate-blogs-front-page-structure.
11. Lubbers P., Albers B., Salim F. Pro HTML5 Programming: Powerful APIs for Richer Internet Application Development. – Berkeley, CA: Apress, 2010. – 304 p.
12. Instapage, https://instapage.com/what-is-conversion-rate-optimization-chapter-8.
13. Ruban, K. Smelyakov,V Martovytskyi, D. Pribylnov and N. Bolohova Method of neural network recognition of ground-based air objects // IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT), 24-27 May 2018. – P. 589-592. DOI: 10.1109/DESSERT.2018.8409200
14. K. Smelyakov, D. Pribylnov, V. Martovytskyi, A. Chupryna Investigation of network infrastructure control parameters for effective intellectual analysis // IEEE 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET), 20-24 Feb. 2018. – P. 983-986. DOI: 10.1109/TCSET.2018.8336359
15. K. Smelyakov, A. Chupryna, D. Yeremenko, A. Sakhon, V. Polezhai Braille Character Recognition Based on Neural Networks // IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 21-25 August 2018. – P. 509-513.
16. G. Churyumov, V. Tokarev, V. Tkachov and S. Partyka, "Scenario of Interaction of the Mobile Technical Objects in the Process of Transmission of Data Streams in Conditions of Impacting the Powerful Electromagnetic Field", 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 2018. – DOI: 10.1109/DSMP.2018.8478539.
17. S. Mashtalir, O. Mikhnova, M. Stolbovyi Sequence Matching for Content-Based Video Retrieval // IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 21-25 August 2018. – P. 549-553.

METHOD FOR ENSURING SURVIVABILITY OF FLYING AD-HOC NETWORK BASED ON STRUCTURAL AND FUNCTIONAL RECONFIGURATION

Genadiy Churyumov, Vitalii Tkachov, Volodymyr Tokariev, Vladyslav Diachenko

Kharkiv National University of Radio Electronics,
61166, Ukraine, Kharkiv, Nauky Ave, 14,
g.churyumov@ieee.org, tkachov@ieee.org, tokarev@ieee.org,
vladyslav.diachenko@nure.ua

Self-organizing flying ad-hoc networks are increasingly used to solve the tasks of recording, storing and transmitting data in space. In some cases, such networks are implemented on a hierarchical basis (mesh topology) and are specialized. These networks are rather volatile since the failure of one of the nodes can disrupt the entire network, and the time it takes to reconfigure the network may be too long. The survivability of the network is an important aspect of the main function (goal achievement) performance reliability. The network has to perform the main function throughout the operating time. During and after the impact of adverse factors, in order to perform the main function, the network has to restore its functions in a minimum time. A characteristic property of a natural or techno-productive negative influence is its low predictability, suddenness, instantaneous distribution, the chance of damaging network nodes, low probability of failure of the network nodes outside the scope of the factor. The goal of survivability is to perform a target function in the event of a malfunction or network failure and the possibility of complete timely recovery in the case of a failure. The article is devoted to the development of a method for ensuring the survivability of flying ad-hoc network. Effective ways to ensure the survivability of the network in adverse conditions is the application of reconfiguration scenarios, redistribution of functions in the network among nodes, temporary self-isolation of nodes, etc. The proposed method is based on the use of the strategy of structural and functional network reconfiguration. This strategy is based on the aggregate-decomposition approach to network nodes. Experimental studies show that the probability of

maintaining the functionality of the network when using the strategy of structural and functional reconfiguration increases the probability of performing the main function during the influence of the negative factor up to 15% and after it - up to 45%. The analysis of the obtained results shows that additional experimental research is needed to accumulate statistical information for modification of the method in the context of the introduction of self-learning elements.

Key words: *survivability, flying ad-hoc network, reconfiguration, data transmission.*

Introduction

Lately most innovations have been appearing in high-tech areas such as biomedical engineering, robotics, infocommunication technologies and artificial intelligence systems. This is quite a logical movement, since many points of contact of directions generate new vectors in the development of science and technology, which led to the massive use of embedded systems, which are the main component of information systems. This is the result of combining technologies of different directions [1]. Demand for the embedded systems is steadily increasing, and together with this, requirements for products on their base are also growing.

The high complexity and rapid development of the elemental base of the embedded systems leads to an increase in the level of abstraction, at which most of design decisions are taken. This requires extensive use of simulation, methods for mathematical analysis and formal verification of models of embedded systems.

An example of the above-mentioned systems is a self-organizing network based on the unmanned aerial vehicle known as a flying ad-hoc network (FANET). Many works [2-5] are devoted to the subject of FANETs, which give a detailed description of the basic principles of design, development and operation of FANETs. Fig. 1 shows one of the possible options for implementing systems, subsystems, classes, types, and components. It should be noted that the problem of external negative influence on the hardware component of the embedded system is considered in this paper. Problems of software deflection due to existing information "viruses" are not considered in this paper. The problem associated with the powerful external influence of microwave radiation on

the hardware component of the embedded system or radio suppression of FANET nodes.

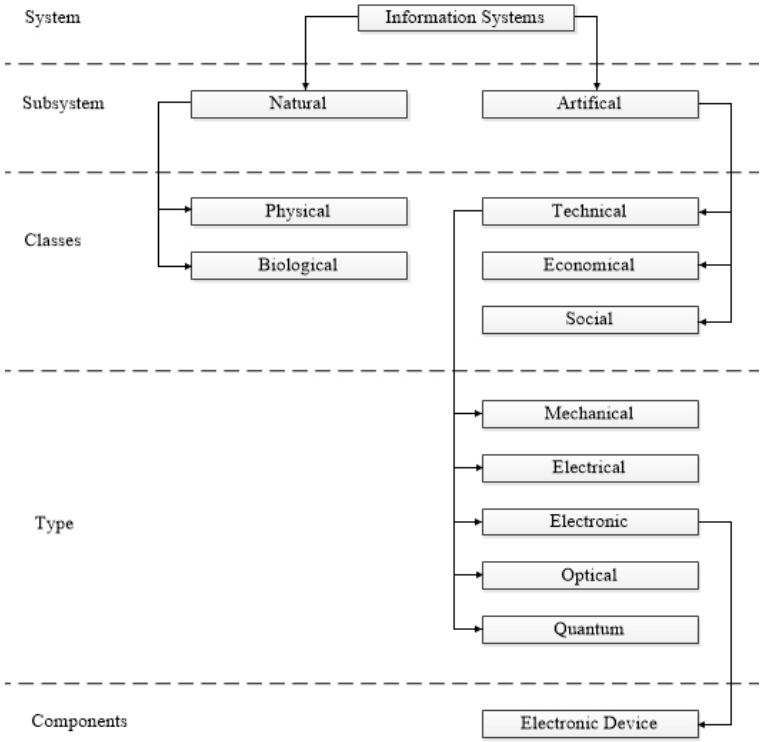


Fig. 1 Scheme of the IS structure

The study of FANET operation is directly related to the structural dynamics of various nature caused by changes in the parameters and state of nodes (UAVs) of the network at different stages of their life cycle under the influence of objective and subjective factors [6]. The natural hazards (lightning, temperature anomalies) as well as technical and industrial activities (electromagnetic pollution, damage), which lead to critical situations and failure of the network in general, represent a particular danger to the functioning of the FANET. In these conditions, one of the most important strategic directions for the development of embedded systems is ensuring the continuity of technological processes of the FANET and increasing functional resistance to failures in the network.

Such an option for managing the structure of objects as a structural and functional reconfiguration of an existing topology of the FANET has become widespread in practice when solving the problems of ensuring the survivability of embedded systems in the theory of structural dynamics management [7].

The reconfiguration of the FANET topology means the process of changing its structure in order to preserve and further restore (increase) the level of network performance or to ensure a minimum reduction in the network efficiency under degradation of its functions [8]. At the same time, the survivability of the system means its ability to adapt to new unforeseen conditions of operation, withstand unwanted external influences while realizing the main function [9].

This work is devoted to the development of the FANET survivability ensurance in adverse conditions, which is based on the structural and functional reconfiguration of the FANET topology.

1 Analysis of known solutions

In the article "Routing protocols in wireless networks" [10], the authors focus on examples of implementing reactive, proactive, and hybrid protocols in solving the problem of optimal construction of data transmission paths between FANET nodes as well as to ensure survivability of the network as a whole. The problem of influence of the dynamics in distance change between the nodes during FANET movement in space (including nodes-participants of the network) on the reconfiguration of the FANET topology is considered from the perspective of searching for optimal data transmission routes. It is argued that in the case of using proactive protocols, there is a need for the regular transmission of service information between FANET nodes to update routing tables and to continuously change the role of nodes. The authors emphasize that in quite dynamic FANETs, there is a problem of buffer overflow for the parameter of sequence numbers of network topologies, the time of their search increases. This is critical for a FANET with fast-changing topology. When using reactive protocols, such as Dynamic Source Routing (DSR), there may be an unreasonable increase in the size of data packet for long routes or an increase in a new address format (IDs), as in IPv6. It is also noted that the general problem of hybrid protocols and geo-routing protocols lies in the narrow specialization and complexity of their implementation [11, 12].

Based on suggestions of the authors on the possibility of a rotary choice of solutions for improving the survivability of a specialized FANET, other shortcomings can be distinguished. First, the FANET can be operated using protocol support within a single protocol family. Creating a FANET that could use a variety of tools in different protocols requires the creation of a complex statistical apparatus, the mechanisms of semantic analysis and the use of self-learning methods [13]. Complication of the system leads to a significant increase in the reaction time on factors, the probability of false positives also increases.

Secondly, each of the protocol types has its disadvantages under conditions of different densities and speeds of nodes. For example, proactive protocols are characterized by advantages over reactive ones in time of rerouting in case of a node failure. In proactive protocols, this process takes place in advance, taking into account prognostic models - it is only necessary to read the scenario of the route from the table, while reactive protocols need to send a broadcast request and expect a response from the recipients. Permanent broadcasting reduces the bandwidth of the FANET for useful data transmission. In addition, hybrid protocols can significantly reduce routing efficiency due to network clustering.

Thirdly, there are no algorithms for network operation in the case of instant degradation or radio suppression of nodes in the FANET in the epicenter of the negative external influence factor. Since the time of the negative external factor influence is less than the time necessary for an adequate response of the FANET to make a decision on the reorganization of the topology or the change of the functioning protocols in order to ensure the survivability of the network, the protocols described in the work, in fact, are meaningless, and achievement of the target function is in jeopardy.

In the article "Viability of wireless communication networks in conditions of emergency sutation" [14], the authors substantiate the thesis that the most effective way to ensure the functioning of amobile communication network in the conditions of the adverse factors is to increase the intensity of service by operating nodes, i.e. searching of functions of the damaged non-operaring nodes by working nodes; increasing the number of communication channels in a damaged cluster of nodes is proposed to be carried out at the expense of redistribution of the released radio frequency resource; transition to the use of other frequency channels. In the final part, the authors conclude that the proposed solution can be used as an additional measure when the FANET falls into the zone of negative external influence of the damage factors.

The disadvantage of this method is the emergence of a significant problem, which consists in solving the problem of eliminating interference of radio signals.

In the paper "Peer selection algorithm in flying ad hoc networks" [15], the authors propose a comprehensive solution to ensure the survivability of the UAV self-organizing network at the expense of: a data transfer method at the application level of the OSI model; an algorithm for selective retransmission request at the application level of the OSI model (AL-ARQ); an algorithm for route selection; an algorithm for choosing an assistant node using the "greedy" criterion; a criterion based on the relative speeds of the nodes. The solution is proposed for the use in the FANET under conditions of radio interference and possible external factors of node damage.

This solution is effective for highly specialized tasks of guaranteed data streaming. Multipath redundancy, permanent reconfiguration of the topology can be realized on the FANET by distributing identical data and forecasting models, collecting and processing static information in various ways. The proposed network encoding also makes the entire process secure. However, the effectiveness of the proposed solution in case of changing the network environment requires additional studies.

2 Rationale for application of FANET topology structural and functional reconfiguration

The structural and functional reconfiguration of the FANET is aimed at changing the network topology and performance characteristics of its technical and organizational subsystem to eliminate the effects of various destructive influences and should take into account the possibility of flexible redistribution of a function, a task and a goal performed by the FANET among the valid nodes with taking into account admissible functioning of the FANET with the worst quality indicators within the allowed limits. During the reconfiguration of topology, the FANET may be located in one of the states $G = \{G_v, v=1, 2, ..., m\}$. Changing states may be caused not only by failures of certain nodes or communication channels, but also by critical situations when only one node of the FANET can remain functioning. For a formal description of possible situations, let us consider some assumptions:

- The feature of the problem statement of structural and functional reconfiguration to ensure the survivability of the FANET is connected to the fact that a set of partial solutions $Q(G_v) = (Q_1(G_v), Q_1(G_v), ..., Q_n(G_v),$

..., $Q_N(G_v)$), $n \in \hat{N} = \{1, 2, \ldots N\}$ of the FANET performance quality of service in the state of G_v can be divided into two groups of indicators according to the following scheme (Figure 2).

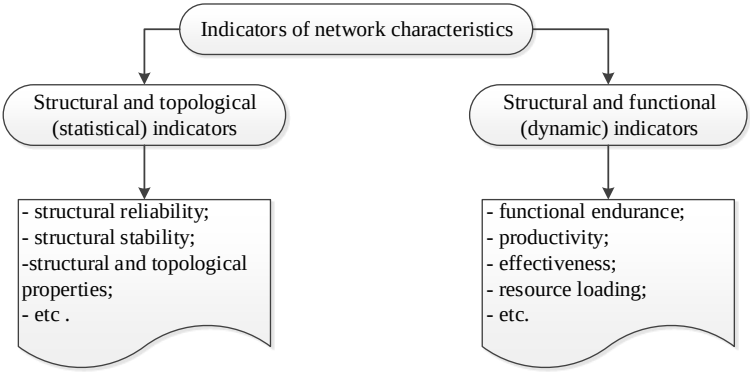


Fig.2. Schematic diagram of FANET characteristics.

The regularity between groups can be described as $\hat{N}_{str} \cup \hat{N}_{fun} = \hat{N}$ under $Q_n(G_v)$, $n \in \hat{N}_{str} \subseteq \hat{N}$, $n \in \hat{N}_{fun} \subseteq \hat{N}$;

- The FANET operation in each of the G_v states is determined by the set of operating and non-operating functional nodes. We will consider those nodes that are not able to perform operations of storing, receiving, transmitting, processing and protecting information resources as non-operating; the nodes performing at least one of these operations will be considered as partially operating;
- structural and functional reconfiguration of the FANET topology occurs under the assumption that a critical situation, unlike a failure (possible or predicted event), is an event that is possible but not probable or its probability is very small and can not be reasonably evaluated during the design of FANET. In other words, the reasons for the emergence of critical situations, as a rule, do not obey probabilistic statistical laws and have a multi-aspect and multifactorial nature;
- analysis of the structural dynamics of the FANET shows that its structures do not change continuously under the influence of various causes, but maintain the stability of the topology at certain time intervals.

Obviously, the value of the partial parameters $Q_n(G_v)$, $n = \{1, 2, \ldots N\}$ as a function of the FANET in each of the G_v states depends on the set of non-operating, partially operating and operating nodes; distribution of operations of processing, storage, reception and transmission of information; redistribution of these operations between the FANET operating or partially operating nodes.

Proceeding from the theory of system survivability, one of the objectives of managing the structural dynamics of the FANET is to provide the maximum possible performance level of the network and its nodes at every moment of time. This goal is achieved by targeted external influence on the degradation process of the FANET in such a way as to eliminate or reduce the possibility (probability) of FANET transitions to the unwanted state, and to manage the processes for updating the FANET.

An important condition for developing a FANET survivability method is to analyze and evaluate its topology. To do this, the theory of structures taxonomy can be applied based on such concepts of homogeneity, equality and monotony [7]. In this approach, it is assumed that the topology of the network is homogeneous if all the functional nodes included in it are identical; and heterogeneous if at least one of its node is different from all others. A FANET structure is considered to be equival if the loss of one of the nodes results in an equal significant loss of any other, and vice versa, the structure is unequal if the individual nodes of the FANET are of great value compared to others. Considering this property, we must further investigate the criticality of input nodes by their functional features. Detection of critical elements contributes to the optimization of the functional policies of other nodes that play the key role in ensuring the reliability, security and survivability of the FANET. The criticality of node failure is considered as a complex property, for the evaluation of which it is expedient to use such partial quality indicators as: failure probability; severity of failure consequences; stability of a node to the influence of external adverse factors; reservation ability; cluster rebuilding; possibility to control node state; duration of failure risk existence; node self-isolation; ability to localize a failure.

This analysis has shown that the model of the FANET functioning can be represented by a structural scheme, a fault and event tree, a connectivity graph, etc. But such a model can describe functioning of a monotone network only. In monotone models, it is impossible to take into account conflicting relationships and relationships between functional nodes: for example, in some configurations, such connections increase the efficiency of the FANET operation while in others they decrease. Such a model can not be operated if there are nodes, which simulnataneously

increase, for example, reliability or security, and the others are the cause of failures or critical situations, that is, they have the opposite, harmful effect on the security of the FANET as a whole.

3 Algorithmic support

Figure 3 shows a general view of the algorithm that implements the above principles. It is important to emphasize that there are a number of additional steps that are not given in Figure 3. These, in particular, are: research tasks on monotony, homogeneity, equivality of FANET structures based on the policy of functional definitions of the basic configuration; assessment of node failure criticality; parametric synthesis of the initial structural architecture of the FANET; multicriteria analysis of node failure criticality; analytical and simulation modeling of the conditions of structural and functional reconfiguration; constructing classes of equival structural reconfiguration scenarios and isolating reference scenarios.

4 Simulation modeling

In the simulation modeling, authors-initiated scenarios for the operation of a group of mobile objects [6, 16] and known software models [17] are used, taking into account the developed algorithmic design.

Let one of the FANET nodes B be connected to the control node A, to which all the collected data are sent. Control node A carries out a control over the FANET and monitors the state of its nodes. The interaction of the FANET nodes is based on the mesh principle [18, 19]. The FANET continuously transmits a data flow. The most stringent requirements for the quality of service, equipment and parameters of the FANET are renderd by nonelastic kind of data. Therefore, in order to improve the efficiency of the FANET, it is advisable to take into account the features of such data.

The model uses physical 802.11 technology at physical and channel levels. The Yans software package is applied as a simulator, which uses the Monte Carlo method. One of the factors that determines the degree to which the FANET simulation model is based on mobile nodes is the choice of an adequate model for location of mobile network devices. In the process of choosing a mobility model, the following points are usually taken into account: the desire to adequately consider the features of the movement of nodes from the point of view of their impact on the aspects of traffic transmission in the network; the need to consider resource constraints and the impact of the detailed description of the movement on

the complication of the FANET model in general; the ability to take into account the requirements for reconfiguring the FANET topology in case of a failure of the nodes with the help of the chosen system of parameters. In the framework of this work, the well-known Gaussian-Markov model of nodes mobility is used [17].

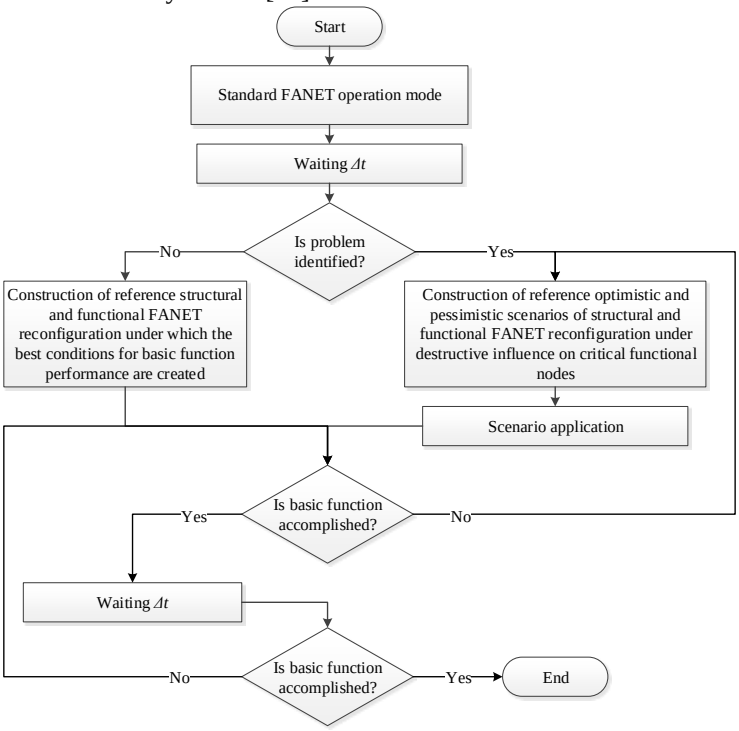


Fig.3. Block diagram of supporting algorithm

This model is the one with memory, that is, the current position of the node in it takes into account its position in the previous step. The movement of mobile nodes in a model is limited to the zone of action of the coordinating node A, where the node changes the direction of its movement after reaching its boundaries. The advantage of the model is the formation of movement trajectories smoothed in speeds and directions, the ability to vary the parameters of movement and the degree of non-determinism of the model, for example, for optimal accounting the influence of external factors that cause deviation of the node from the calculated lanes.

The ns-3 simulation system has been chosen as the basic tool for simulation modeling [17]. The total time of FANET simulation is 200 s. The transfer of useful data and the exchange of official information between nodes begins in the interval between the 10th and 11th seconds and lasts until the end of the simulation. Traffic transmission is simulated by the flow of JUDP datagrams at the data transfer rate of 1024 Mbit/s. To simulate the flow at the output of the node B, the sequence of packets of fixed length is set to 1024 bits. When simulating, nodes move randomly in the area of 500x500 m; the number of nodes is 11 (1 control node (A), 1 node (B), 3 relay nodes - (C), 6 data logger nodes (D)); node movement speed - up to 20 m/s; transmitter power - 8 dBm; routing protocol - AODV. The basic FANET topology is shown in Figure 4. At the 25th and 100th seconds in the FANET two nodes are lost: the relay node and the data logger node.

As a result of the experiment, the dependence of the intensity of data entry on node A on time (Fig. 5) was obtained. Descriptive part of events in different time intervals is shown in Table. 1

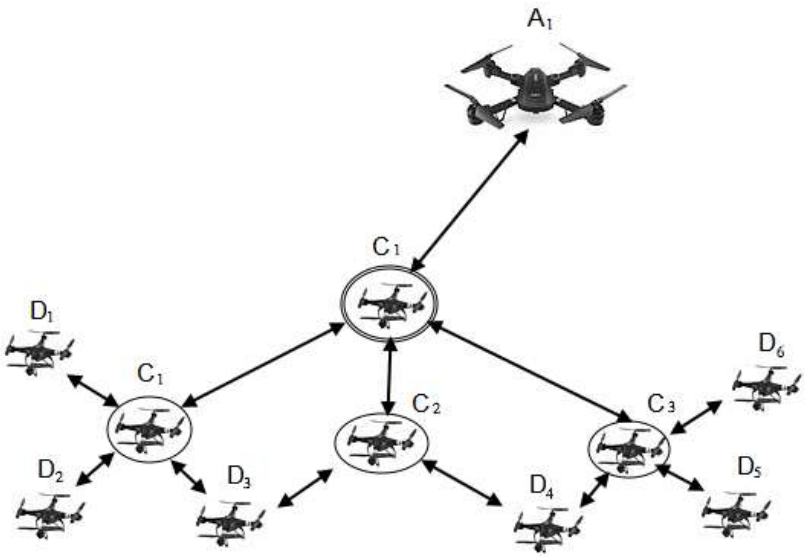


Fig.4. The basic topology of the FANET under study.

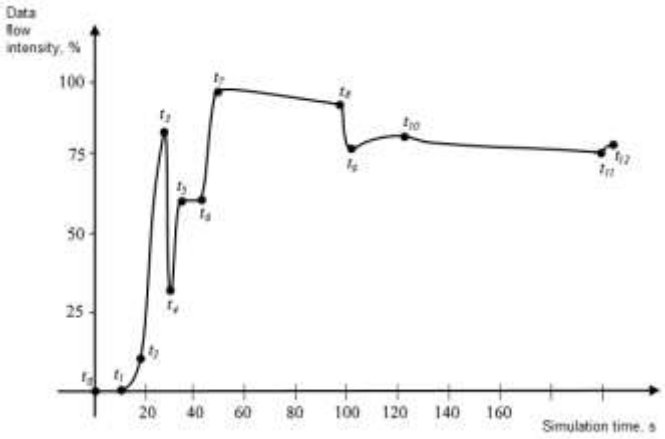


Fig.5. Structural diagram of network characteristics.

Table 3. Events in the FANET during the experiment

Time range		Event
Δt_1	0-10 s	Deploying the FANET.
Δt_2	11-18 s	Setting up the basic FANET configuration, constructing a routing table.
Δt_3	19-25 s	Data transmission.
Δt_4	25-30 s	Failure of node the C_3. FANET assessment of a situation.
Δt_5	30-34 s	Reconfiguration of data flows between the nodes D_4–C_2.
Δt_6	35-42 s	Movement of node the C_2 to the point of the former node C_3
Δt_7	43-50 s	Data transmission under the new configuration.
Δt_8	51-100 s	Reducing the intensity of the flow of service data.
Δt_9	101-102 s	Failure of the node D_6. FANET assessment of a situation.
Δt_{10}	103-121 s	Change location of the nodes D_4 and D_5.
Δt_{11}	122-187 s	Data transmission under the new configuration. Reducing the intensity of the service data flow.
Δt_{12}	187-200 s	Broadcast message about the return of nodes to the base.

Structural and functional reconfiguration, according to the algorithmic description, should take several seconds, however, the abnormal failure rates of the packet transmission on the curve lasted up to 4 seconds (Δt_5).

At node speeds of about 10 m/s, this is explained probably due to the inability to instantly reach the required position, taking into account the density and mobility of the nodes, or the long connection when the node-data logger is released from the zone of direct reach to the relay node.

Based on the graph (Figure 4), we can conclude that the use of structural and functional reconfiguration, in general, leads to a significant improvement in the quality of transfer of non-elastic traffic from the nodes of the FANET.

It should be noted that in this experiment, even with the highest value of the number of nodes, the percentage of service traffic is less than 5%. This may be, first of all, due to the fact that the FANET model developed in this study is rather simplified compared to the actual FANET and does not consider the presence of "background" data from other nodes as well as service information that is not related to routing.

5 Conclusions and recommendations

The analysis of the FANET topology reconfiguration to ensure its survivability is relevant nowadays. Existing formulations of the task of reconfiguration are characterized by high dimensionality and do not take into account most of the operations. In order to consider the features of the FANET management, general and partial requirements for the development of new principles, models, methods and techniques of multicriteria assessment, analysis and selection of structural and functional reconfiguration of the FANET topology are formulated and substantiated. The analysis of these requirements allowed formulating the direction of the aggregate-decomposition approach to solving the problem of structural and functional reconfiguration of the FANET topology.

The simulation results of FANET showed that the developed method is working and provides data transmission in case of a failure of network nodes. At the same time, there is a need to refine algorithmic supporting in order to approximate it from theoretical calculations to the real FANET. This can be achieved by entering service data from other nodes, accounting for non-routing service traffic, applying data of Big Data class, etc [20]. As a mobility model, it is recommended to use a more organized algorithm for nodes. For example, enter a task for nodes B to correct the position of nodes D. It is desirable to introduce routing protocols more adapted to FANETs into the simulation model.

The research is carried out within the framework of the implementation of the fundamental research work "Creation of Scientific and Methodological Foundations for Ensuring Survivability of Network

Information Exchange Systems in Conditions of External Influence of High-frequency Microwave Radiation" on the basis of the educational and scientific Laboratory of Reconfigurable and Mobile Systems of the Department of Electronic Computers in Kharkov National University of Radio Electronics.

Reference

1.T. Kuhn The structure of scientific revolutions. Chicago, Ill.: The University of Chicago Press, 2015.

2.Singh K., Verma A. K. Flying Ad hoc Networks Concept and Challenges //Encyclopedia of Information Science and Technology, Fourth Edition, IGI Global, 2018, pp. 6106-6113.

3.Mukherjee A. et al. Flying Ad hoc Networks: A Comprehensive Survey // Information and Decision Sciences, Springer, Singapore, 2018, pp. 569-580.

4.İ. Bekmezci, O. Sahingoz and Ş. Temel, "Flying Ad-Hoc Networks (FANETs): A survey", Ad Hoc Networks, vol. 11, no. 3, pp. 1254-1270, 2013. – DOI: 10.1109/EDM.2018.8434973.

5.A. Leonov and G. Litvinov, "Simulation-Based Packet Delivery Performance Evaluation with Different Parameters in Flying Ad-Hoc Network (FANET) using AODV and OLSR", Journal of Physics: Conference Series, vol. 1015, p. 032178, 2018. – DOI: 10.1109/EDM.2018.8434973.

6.I. V. Ruban, G. I. Churyumov, V. V. Tokarev, V. M. Tkachov, "Provision of Survivability of Reconfigurable Mobile System on Exposure to High-Power Electromagnetic Radiation", Selected Papers of the XVII International Scientific and Practical Conference on Information Technologies and Security (ITS 2017), CEUR Workshop Processing, pp. 105-111, November 30, 2017.

7.Pavlov A. N. Metodologicheskiye osnovy resheniya problemy planirovaniya struktur-no-funktsional'noy rekonfiguratsii slozhnykh ob"yektov // Izvestiya vysshikh uchebnykh zavedeniy. Priborostroyeniye. – 2012. – T. 55. – №. 11. – Pp. 7-13.

8.Dodonov A.G., Kuznetsova M.G., Gorbachik Ye.S. Vvedeniye v teoriyu zhivuchesti vychis-litel'nykh sistem. – K.: Nauk. dumka, 1990. – 184 s.

9.A.G. Dodonov, D.V. Lande Zhivuchest' informatsionnykh sistem. – K.: Nauk. dumka, 2011. – 256 s.

10. Mehta, Komal, and Raju Pal. "Energy Efficient Routing Protocols for Wireless Sensor Networks: A Survey." Energy 165.3 (2017).

11. Yeremenko O., Lemeshko O., Persikov A. Secure Routing in Reliable Networks: Proactive and Reactive Approach. Advances in Intelligent Systems and Computing II, CSIT 2017, Advances in Intelligent Systems and Computing, Springer, Cham. 2018. Vol. 689. P. 631–655.

12. Lemeshko O., Yeremenko O. Enhanced method of fast re-routing with load balancing in software-defined networks. Journal of ELECTRICAL ENGINEERING. 2017. Vol. 68, Issue 6. P. 444–454.

13. Ruban, K. Smelyakov,V Martovytskyi, D. Pribylnov and N. Bolohova Method of neural network recognition of ground-based air objects // IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT), 24-27 May 2018. – P. 589-592. DOI: 10.1109/DESSERT.2018.8409200

14. Romashkova O.N. Zhivuchest' besprovodnykh setey svyazi v usloviyakh chrezvychaynoy si-tuatsii / O.N. Romashkova, Ye.V. Dedova // T-Comm: Telekommunikatsii i Transport, 2014. – №6. – S. 40-43.

15. D. Vasiliev, A. Abilov and V. Khvorenkov, "Peer selection algorithm in flying ad hoc networks", 2016 International Siberian Conference on Control and Communications (SIBCON), 2016. – DOI: 10.1109/SIBCON.2016.7491734.

16. G. Churyumov, V. Tokarev, V. Tkachov and S. Partyka, "Scenario of Interaction of the Mobile Technical Objects in the Process of Transmission of Data Streams in Conditions of Impacting the Powerful Electromagnetic Field", 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 2018.

17. Dorokhova A. A., Paramonov A. I. Issledovaniye trafika i kachestva obsluzhivaniya v samoorganizuyushchikhsya setyakh na baze BPLA // Informatsionnyye tekhnologii i telekom-munikatsii. – 2016. – T. 4. – №. 2. – S. 12-25.

18. Lemeshko O., Yeremenko O., Nevzorova O. Hierarchical Method of Inter-Area Fast Rerouting. Transport and Telecommunication Journal. 2017. Vol. 18, Issue 2. P. 155–167. DOI: 10.1515/ttj-2017-0015.

19. Yeremenko O. S., Lemeshko O. V., Nevzorova O. S., Hailan A. M. Method of Hierarchical QoS Routing Based on the Network Resource Reservation. Electrical and Computer Engineering (UKRCON): Proceedings of the First Ukraine Conference, Kiev, Ukraine, 29 May – 2 June, 2017. IEEE, 2017. P. 971–976.

20. K. Smelyakov, D. Pribylnov, V. Martovytskyi, A. Chupryna Investigation of network infrastructure control parameters for effective intellectual analysis // IEEE 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET), 20-24 Feb. 2018. – P. 983-986.

A HYBRID METHOD OF INFORMATION AGGREGATION FOR COMMUNITY-LEVEL DECISION-MAKING

Kadenko S.V.
Institute for Information Recording of the National Academy of Sciences of Ukraine, Kyiv, Ukraine
seriga2009@gmail.com

When it comes to community-level decision making it is appropriate to utilize expert data based-methods, as the respective subject domains are, mostly, weakly structured ones. At the same time, during decision-making, opinion of the target territorial community members should be taken into consideration alongside expert data. The paper outlines an original method for formal description of weakly structured community-level problems, which uses both expert information and opinion of respondents from among community members. It represents a hybrid approach, incorporating elements of both traditional expert data-based methods and social surveys (questionnaires). The main goal (problem) is formulated by a decision-maker or research organizer. It is then decomposed by experts into sub-goals or factors that are crucial for its achievement, and these factors and their weights are estimated by respondents who are ordinary community members. The method includes the following conceptual steps: hierarchical decomposition of the problem, direct estimation of importance of factors that influence the problem, estimation of lowest-level "non-decomposable" factors by respondents in Likert agreement scale, and rating of the factors based on respondents' estimates through linear convolution (weighted summing). The obtained ratings provide the basis for defining top-priority activities that should be performed in order to solve the problem, and for subsequent distribution of limited resources among these activities. Experimental results, obtained in the process of actual research of public space quality, illustrate the method's application, and confirm its high efficiency. The advantages of the suggested method are efficiency and, at the same time, ease of use. In contrast to traditional expert data-based methods, it does not require any preliminary coaching sessions to be held with the respondents. The method is intended for decision-making support at the level of territorial communities (urban, rural, district, neighborhood, and others) in the spheres, directly related to the interests of community members. Target users of the method include local self-government bodies, media, public and volunteer organizations, activists, and other interested parties..

Keywords: *information aggregation, decision-making support, weakly structured subject domain, expert estimate, hierarchic problem decomposition, Likert agreement scale.*

Introduction: problem relevance and existing approaches

Weakly structured nature is inherent for many fields of human activity. As we know from numerous sources, the characteristic features of a weakly structured subject domain are as follows (see, for instance, [1]): it is problematic to provide a formal description and build analytical models; there are no benchmarks; all decisions made are unique ones; decision-making space dimensionality is very large; the domain is influenced by multiple significant criteria; information on the objects is incomplete, and human factor plays a considerable role.

At the same time, in order for decisions made in any subject domain (a weakly-structured one as well) to be efficient, they should be informed and well-substantiated. Thanks to consideration and systematization of all available information, the level of trustworthiness of the decision-maker (DM) increases, while the possibility of erroneous and incompetent decisions being made is reduced. So, in order to set priorities and plan respective activities, a DM needs to be able to analyze and formally describe weakly structured subject domains. Consequently, the problem of analysis and formal description of these domains retains its high relevance.

With the listed properties of weakly structured subject domains in mind, we should acknowledge that expert data-based methods and technologies are a powerful (and often, the only) mathematical tool for decision-making support in these domains. Experts (ideally – narrow-profile specialists) should be engaged, first, to outline a set of factors which are crucial for a given domain, and identify the nature of interrelations between these factors, and, second, to provide numeric (quantitative, cardinal), or at least, ordinal (rank) estimates of the relative weights of factors and alternative decision variants, from which the DM will have to select an optimal one.

In the most common case, decision-making means either choosing one of several alternative decision variants from a given set, or ranking/rating of these variants. Optimality of a decision variant (alternative) is defined based on some specific global (aggregate) efficiency criterion, which can reflect the degree of achievement of a certain main goal. Such general efficiency criterion is, usually, defined based on the realities of a specific situation. It provides the “starting point” in the process of formal analysis and decomposition (break-down) of the subject domain into particular aspects, as well as expert estimation of decision variants. In a way, the

global efficiency criterion plays the role of a target function from mathematical optimization theory [2], or a utility function from utility theory [3]. However, we should stress that one of the peculiar features of weakly structured subject domains is the impossibility of analytic expression of this function. In fact, the experts are involved in order to define its specific (non-analytic!) look and an optimal decision variant (in accordance to a given efficiency criterion).

The result of expert decomposition of the main criterion (goal) into sub-criteria (sub-goals) is a hierarchy of criteria, which characterize the subject domain.

Expert data-based decision-support methods are listed and described in multiple publications. We can mention classical works by Kendall [4], Arrow [5], Kemeny [6], Fishburn [7], Saaty [8] and others. Soviet and Ukrainian authors (including Mirkin [9], Litvak [10], Totsenko [11], Gnatiienko and Snytiuk [12], and others) also largely contributed to development of these methods. World-known multi-criteria decision-making methods include AHP/ANP [8], TOPSIS [13], ELECTRE [14], and others. In present-day Ukraine popular decision support methods include technological forecasting [15], complex target-oriented dynamic estimation of alternatives (CTDEA) [11, 16], and various interval estimation methods [12].

Acknowledged multi-criteria expert estimation methods include the aforementioned AHP, TOPSIS, ELECTRE, CTDEA methods; the most popular ranking aggregation methods include Borda rule, Condorced rule, Kemeny's median, and others; and when it comes to pair-wise comparisons, the common approaches include, again, AHP/ANP [8], "line", "triangle", "square" methods [11], and combinatorial approach [17, 18].

Specific applications of expert data-based methods, particularly those using the hierarchical approach to problem decomposition, are rather numerous. Dozens of specific applications are described in the proceedings of the International Symposiums for the Analytic Hierarchy Process (ISAHP) [19].

In modern-time Ukraine, there are several spheres, calling for expert data usage in decision-making process. In this context we can, again, mention the technological forecasting problems [15], socio-economic development planning [20], environmental protection [21], and other spheres. Warfare [22, 23] and information security and related decisions

have gained relevance for Ukraine in recent years. Weakly-structured nature of these spheres is demonstrated in [24–26].

Another subject domain, which is relevant for Ukraine, is decision-making at the level of communities (village, raion, urban, territorial, etc). After launching of decentralization (particularly, budget decentralization), administrative-territorial organization reform, and formation of unified territorial communities (see [27, 28]) the role of communities in decision-making substantially increased. Consequently, there is a need for efficient yet easy-to-use analytic tools, which would provide an opportunity to consider the opinion of community representatives during decision-making. Only when public opinion is taken into account, community-level decisions will truly reflect the interests of community members. Specific decisions, immediately concerning community members, are related to such issues as planning and improvement of transport and road networks, domestic waste disposal, water supply and disposal, gentrification and improvement of territories, construction of recreation zones, reintegration of public usage locations (museums, libraries, parks, etc) into community life etc.

While the arsenal of available approaches and methods is seemingly wide, in this particular case we are talking about a specific type of problems and specific conditions of expert examination. So, the question is: which of the listed approaches (or their components) should be applied in community-level decision-making to ensure that community members' opinion is taken into account?

1 In Description of problem class and solution idea

The key feature of the aforementioned class of problems is their weakly structured nature. Formal problem statement is possible only when the goal, which the DM or other interested party is trying to achieve, is clearly defined. A community often finds it hard even to formulate a specific problem, not to mention identification of factors, which could influence its solution.

"What should be done with the waterfront area?", "how can we reorganize the park?", "what is the best way to arrange the system of water supply and water disposal in the settlement?", "what should we do with an old village cultural center (club), museum, library?" etc. Such typical community-level problem examples have several features in common. First, as it has been said, they immediately concern the

representatives of a given community and reflect their interests. Second, they do not include any recommendations as to how the issue under consideration can be resolved. That is, the only input data is some problem to be solved, or some main goal to be achieved (let us denote it as G, and let us denote the main criterion of efficiency of its achievement as C_0).

In view of weakly structured nature of the problems, it would be reasonable to hierarchically decompose the problem into specific components. Relevance and effectiveness of application of hierarchical approaches are shown, for example, in the works of Saaty ([8]), Totsenko ([11]), Gnatiienko and Snytiuk ([12]), Pankratova and Nedashkovskaya [29, 30], and others. The hierarchical approach proved to be an effective instrument in a multitude of applications (see [19]).

So, at the initial phase it is suggested to decompose (break-down) the main goal or problem into factors, which play decisive roles in its achievement (solution), as it is done in the listed methods, such as AHP [8], TOPSIS [13], or CTDEA [11, 16].

We should remember that formulations of criteria should be easy-to-understand, while criteria hierarchy graph should be balanced and not overloaded with excessive number of connections (edges) and nodes (more detailed requirements to the process of hierarchy building are described in [1, 31]).

Both in AHP and CTDEA, when a hierarchy is built, the weights of impact factors (criteria) are defined, that is every edge of the hierarchy graph is assigned a certain weight.

In [1, 31] it was stressed that the scale used for estimation of criteria had to be understandable for an unprepared expert or respondent. In order to achieve this, we should choose the scale with grades described by verbal rather than numeric values. Beside that, the scale should not make the respondent keep too many values and objects in mind simultaneously. For example, in a decision support technology, described in [32], the expert has to select the type of an ordinal pair-wise comparison ("more-less"), number of scale grades, and a particular grade from the chosen scale. This process calls for preliminary coaching sessions to be held with experts.

In order for the opinion of community members to be taken into account, they should be involved in the process of formal description of a given weakly structured problem. Moreover, the DM (or analytic research organizer) should keep in mind that, in the general case, it is impossible to

organize coaching sessions with all the respondents. So, the process of hierarchy building, particular look of the graph, specific criterion formulations, and the scale, in which the importance of criteria is estimated, should be as easy-to-understand as possible.

Based on these considerations, it is suggested to delegate the hierarchy building process to the experts in the given subject domain (as it is done in AHP or CTDEA), while introducing several additional requirements:

1) We should forbid input of cycles (loops, where criteria influence themselves) into the hierarchy graph. In the ideal case the hierarchy should represent a tree-type graph, that is, one node should have only one "ancestor" (hierarchy graphs of this type are addressed, for example, in [33, 34]).

2) Bottom-level criteria should be formulated not as concepts (for example, "quality of family leisure in the park, estimated in the scale from 1 to *n*"), but as positive statements, with which a respondent (not an expert, in the general case!) might agree or disagree. For example: "the park is a good place to spend quality time with a family"; response variants: "totally agree", "agree", "disagree", "totally disagree", "don't know". That is, we should provide respondents with an opportunity to estimate "atomic" bottom-level factors in Likert's scale [35]. Simplicity and vividness of this approach, as well as numerous sociological studies, in which the approach is successfully used, speak in its favor. Besides that, the choice of Likert's scale results from the need to consider opinions of a large quantity of respondents (and not just of a few experts, as in case of "classical" expert estimation methods).

3) Estimation of relative weights or impacts of factors (criteria) should be delegated to respondents from among target community members, so these estimates should be direct ones. In the process of weight estimation, in order to ensure vividness, the estimates are to be provided in graphic (and not verbal or numeric) format (this requirement is also based on the peculiarities of obtaining data from respondents, described in [1, 31]).

Based on the listed requirements, we can suggest the following step-by-step algorithm of formal description of a weakly structured problem.

2 Step-by-step algorithm of problem solution

1) The DM or expert examination organizer formulates the main problem or goal G and chooses the experts (at least one expert) in the respective subject domain.

2) Experts build a hierarchy of criteria (see Fig. 1 below), which are crucial for the given problem or goal: $\{C_i : i = 0..n\}$. The bottom-level criteria (which do not have ancestors in the hierarchy graph) $\{C_{i_k} : k = 1..l\}$ are formulated as positive statements, which the respondents will estimate in Likert's agreement-disagreement scale.

3) A set of respondents $\{r_j : j = 1..m\}$ is chosen from among the members of the given community.

4) Respondents estimate (directly, in clear, preferably, graphic format) the weights of impact factors at each hierarchy level. As a result, we get a set of impact coefficients (or relative weights) $\{w_i^{(j)} : i = 1..n; j = 1..m\}$, provided by every respondent.

Impact w_{i0} of criterion C_i upon the main criterion C_0 is defined as shown in [11] (case of a hierarchy of "tree" or "network" type) and [34], according to formula (1).

$$w_{i0} = \sum_{p=1}^{n_i} \prod_{s=1}^{n_p} w_s \qquad (1)$$

where n_i is the quantity of all possible paths from node C_i to node C_0 in the criteria hierarchy graph; p is the number of a particular path; n_p is the length of the path number p, leading from C_i to C_0; s is the number of the node within the path; w_s is the impact of criterion C_s upon its immediate ancestor in the path number p in the hierarchy graph.

We should note that if the hierarchy graph is a tree (every node (vertex) has no more than one ancestor, as on Fig. 1), the number of summands in formula (1) equals 1, as there is only one path, leading from any criterion C_i to the main criterion C_0.

5) Respondents select the estimates of bottom-level criteria in Likert's scale (as shown above). Scales grades are assigned the respective numeric equivalents for further aggregation, i.e., convolution (for instance, "6 – totally agree", "5 – agree", "4 – rather agree than disagree", "3 – rather disagree than agree", "2 – disagree", "1 – totally disagree", "0 – don't know or don't care"). As a result, we get a certain number of individual judgments of respondents for criteria from the bottom level of

the hierarchy: $\{q_{i_k}^{(j)}; k = 1..l; j = 1..m\}$, where i_k are numbers of the bottom-level criteria, l is their total number (quantity), and j is the respondent's number.

6) Estimates according to every criterion are aggregated through weighted summation (convolution) of estimates, provided by all respondents. As a result, we get the ratings of all bottom-level criteria Q_{i_k};

$$Q_{i_k} = \sum_{j=1}^{m} w_{i_k 0}^{(j)} q_{i_k}^{(j)} \qquad (2)$$

where $w_{i_k 0}^{(j)}$ is the relative impact of bottom-level criterion C_{i_k} upon the main criterion C_0, calculated according to formula (1) based on values of impacts of all intermediate-level criteria, provided by the respondent number j.

Based on these ratings, potential and priorities of the problem solution are identified, while community interests are taken into account. Highest-rated factors represent the top-priority aspects, which the community is satisfied with, while lowest-rated factors represent aspects, which do not satisfy the community, or are insignificant in the eyes of community members.

7) In order to check the consistency of respondents' judgments, we can ask them to write small verbal reviews, in which they should try to outline (once again, this time, verbally) positive and negative aspects, characterizing the subject domain.

We should stress that we are talking about inner consistency of judgments of each respondent: verbal review should be consistent with estimates, provided in Likert's scale at previous steps (see step 5). Mutual incompatibility of the estimates is not a problem, i.e. judgments of different respondents can differ, and it is quite natural, because in our case respondents only express their opinions, and do not try to estimate some objective values.

As judgments are provided in Likert scale, and it is only inner consistency of the estimates that is verified, traditional consistency measures (such as Kendall's rank correlation [4] for ordinal estimates, or consistency index (CI) and ration (CR) [8] for pair-wise comparisons,

spectral consistency coefficient [11], double entropy index [36]) are not applicable. And that is why it makes sense for respondents to write verbal reviews.

8) In order to ensure transparency, tag (word) clouds can be generated (based on verbal reviews), which will also, in a way, represent the ratings of positive and negative aspects of the issue under consideration. In order to build tag clouds based on word frequency analysis of verbal review texts, publicly available online software tools can be used (such as WordItOut, Worlde, WordArt, WordCloud, TagUl, Many Eyes, Tagxedo, etc).

The final result of the algorithm is a formal analytic description of the subject domain. Such a description allows interested parties to clearly identify the key positive and negative aspects in the given subject domain, taking public opinion into account, and, thus, define its potential, prospects, and top-priority problems, at which resources should be targeted.

2 Example

As an example of application of the suggested weakly structured subject domain description method, let us consider an actual research, which took place in summer of 2018, in one of the raion centers of Kyiv oblast (Ukraine). The main task was to study the quality of a public location (the territory of the state local history museum) in order to facilitate its further transformation and improvement. The study envisioned a set of tasks: 1) ensure communication (productive contact) between the community and the town management; 2) identify strong and weak points of the location; 3) identify priorities of the community (i.e. what was important for the residents); 4) evaluate the extent to which the location meets the needs and expectations of its users; 5) stimulate new ideas concerning location improvement; 6) define priorities of location development.

The focus-group from among the community members included 11 respondents, featuring local residents, museum employees, civil servants, journalists, artists.

The main goal was to improve the quality of the location; respectively, the generalized location quality was the main criterion.

Criteria hierarchy was built based on the SpaceShaper [37] methodology guidelines published by the British Commission for

Architecture and Built Environment (CABE). SpaceShaper proved to be an effective, easy-to-use, and affordable instrument of public space improvement, particularly, in the British Commonwealth countries. Particular examples of successful application of the methodology for improvement of various public spaces can be found, for example, in [38–40]. In Ukraine different public organizations presently make the first attempts to use SpaceShaper methodology and its separate components for evaluation of quality and for transformation of public spaces.

The hierarchy of criteria, which influence the quality of the location, built in "Solon" DSS [10, 39], is shown on Fig. 1.

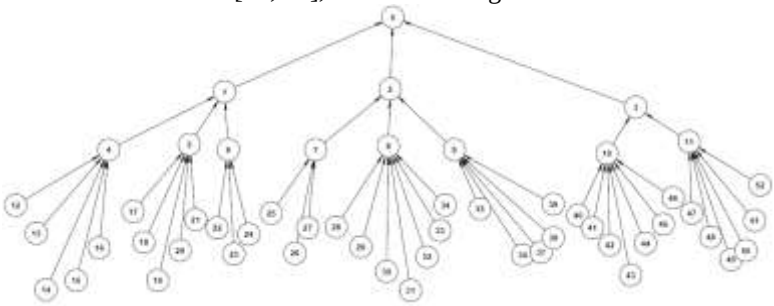


Fig. 1. The look of criteria hierarchy, built in "Solon" DSS

The list of criteria is presented in Table 1.

Table 1. The list of criteria from the hierarchy

#	**Formulation**
0.	Location quality
1.	Functionality
2.	Location characteristics
3.	Location value
4.	Accessibility
5.	Use
6.	Interests of residents (community members)
7.	Order
8.	Environment
9.	Design and look (appearance)
10.	Community (other people)
11.	You (individual respondent)
12.	It is easy to get here
13.	The place is easy-to-find
14.	Orientation is easy
15.	The place is open, whenever I come here

16. I know what is happening here
17. Here I can do what I want
18. The place has everything I need
19. I can enjoy the nature here
20. Here I can learn a lot about local history, flora, fauna, art, etc
21. The place helps me to keep healthy
22. The place is popular among different people
23. A lot of different activities take place here
24. There are no conflicts between different location users
25. It is clean here
26. The place is cared for and looked after
27. People, who take care about the location, are always available
28. Here you can hide from bad weather
29. The place is worth visiting any time of the year
30. The place is well-lighted
31. The air is clean
32. The place is not noisy
33. I feel safe here
34. I am totally satisfied with the location size
35. The place adorns the neighborhood
36. The place is well-equipped
37. You can witness the diversity of plants and animals here
38. The place is inspiring
39. The place is nice
40. The location is an important component of the landscape
41. The locals are involved in location maintenance and organization of events on its territory
42. Everyone feels at home here
43. It is a comfortable place for communication
44. The locals are proud of this place
45. The place is well-located
46. The place is attractive for small business
47. I feel well here
48. I can entertain myself here
49. I can relax here
50. The place is good for thinking
51. I like being here
52. I come here to hide (escape) from routine

As we can see, the topmost (first) level of the hierarchy consists only of the main criterion “the overall quality (efficiency) of the location”. Its immediate sub-criteria (descendants) are functionality (which, in turn, includes accessibility, ease of use, and interests of residents), characteristics (including order, environment, and look/appearance/design), and value of the location (for the community and an individual respondent respectively). The bottom (fourth) level features 41 criterion, formulated as positive statements, with which respondents can agree or disagree (for instance, sub-criteria of

accessibility are: “it is easy to get here”, “the place is open whenever I come here” etc).

Every respondent received a questionnaire form, in which (s)he had to specify his(her) occupation and location usage mode (frequency and purpose of visits), provide weights of criteria of the second and third levels of the hierarchy, and estimate the bottom-level criteria in Likert’s agreement-disagreement scale.

Estimation of weights of intermediate-level criteria was performed using sector diagrams (pie charts) (see Fig. 2). The choice of this particular method for weight estimation results, primarily, from simplicity and transparency considerations. As a result of estimation, every second-level criterion was assigned an integer-value weight from 0 to 12 (sum of all second-level criterion weights equals 12), while every third-level criterion was assigned a weight from 0 to 144 (sum of all third-level criterion weights equals 144). Weights of these criteria were estimated by each respondent. It should be stressed, that this particular method of criterion weight estimation was chosen for the specific study. In the general case, depending on specific hierarchy structure, and the degree of process automation, other weight input methods can be used, providing they are clear and understandable.

Fig. 2. An example of criterion weight input using pie chart

Bottom-level criteria were presented to respondents in the form of tables (in accordance to SpaceShaper methodology, their weights are considered equal). Respondents had to express their judgment on each of the 41 bottom-level criterion in Likert’s scale. An example of the table from the questionnaire, filled out by every respondent, is shown in Table 2.

Once the surveys were completed, the respondents were offered to describe in their own words the strong and weak points of the location, as well as their vision of the ideal condition of the location (so-called "letter from the future"). Aggregation of survey data was performed as follows.

Verbal values were replaced by numeric equivalents (as shown in the previous section): 6 – totally agree", "5 – agree", "4 – rather agree than disagree", "3 – rather disagree than agree", "2 – disagree", "1 – totally disagree", "0 – don't know or don't care".

After that rating of every bottom-level criterion was calculated through weighted summation (convolution) of estimates provided by all respondents (according to formula (2)).

In general case bottom-level ratings lie within the range between 0 and $R_{\max}$:

$$R_{\max k} = mq_{\max} w_{\max k}, \qquad (3)$$

where k is the number of bottom-level criterion; m is the quantity of respondents; $q_{\max}$ is the maximum value of numeric equivalent of a scale grade, which can be chosen by a respondent, while $w_{\max k}$ is a maximum possible criterion weight value.

In the case of our particular problem the number of respondents is $m = 11$; the range of bottom-level criterion weights is $\{w_{i_k} \in Z \cap [0;144]; \sum_{k=1}^{l} w_{i_k} = w_{\max} = 144; l = 41; i_k = \{12..52\}\}$, i.e. bottom-level criteria from Table 1 with numbers 12 to 52 can be re-numbered from 1 to 41; their weights are expressed by integer values from 0 to 144, and their sum equals 144. The range of numeric equivalents of Likert scale grades lies between "0" ("don't know or don't care") and "6" ("totally agree"), i.e in formula (3) $q_{\max} = 6$. Respectively, bottom-level criterion ratings will belong to the range from 0 to 9504 (according to (3) $R_{\max} = 11 \times 6 \times 144 = 9504$). If all ratings need to fall within the range between 0 and 1, they can be normalized.

Table 2 Survey fragment

	Totally agree	Agree	Rather agree	Rather disagree	Disagree	Totally disagree	Don't know	Don't care
Accessibility								
It is easy to get here								
The place is easy-to-find								
Orientation is easy								
The place is open, whenever I come here								
I know what is happening here								
Use								
Here I can do what I want								
The place has everything I need								
I can enjoy the nature here								
Here I can learn a lot about local history, flora, fauna, art etc								
The place helps me to keep healthy								
Interests of residents (community)								
The place is popular among different people								
A lot of different activities take place here								
There are no conflicts between different location users								

Bottom-level criterion ratings, obtained through aggregation of respondents' estimates, are shown on Fig. 3. For the sake of convenience these criteria are numbered from 1 to 41 (as described above).

We should note that absolute rating value does not play any significant role. It is the ratios (or differences) between ratings of different criteria that matter. Besides that, important information can be obtained from the ratios of ratings within each sub-group (functionality, characteristics, value of location). Fig.4 displays the relative ratings of criteria, which belong to "functionality" subgroup.

Similarly, the respective ratings for each intermediate-level criterion were calculated. Rating of a criterion which has descendants in the

hierarchy graph is calculated as the weighted sum of ratings of its immediate sub-criteria:

$$Q_i = \sum_{j=1}^{m} Q_i^{(j)} = \sum_{j=1}^{m} \sum_{k=1}^{v} w_{i_k,i}^{(j)} Q_{i_k}^{(j)} , \qquad (4)$$

where Q_i is the rating of criterion C_i; $Q_i^{(j)}$ is the rating of this criterion calculated based on estimates provided by respondent number j; m is the total number of respondents; v is the number of immediate sub-criteria of C_i in the hierarchy graph; $w_{i_k,i}^{(j)}$ is the weight of impact of sub-criterion number i_k upon criterion C_i, estimated by respondent number j; $Q_{i_k}^{(j)}$ is the rating of the sub-criterion number i_k, calculated based on estimates provided by respondent number j. For instance, in the hierarchy on Fig. 1 the sub-criteria of "Functionality" (C_1) are "Accessibility" (C_4), "Use (C_5), and "Interests of community" (C_6). So, the rating of "Functionality" is calculated as the sum of ratings of these sub-criteria.

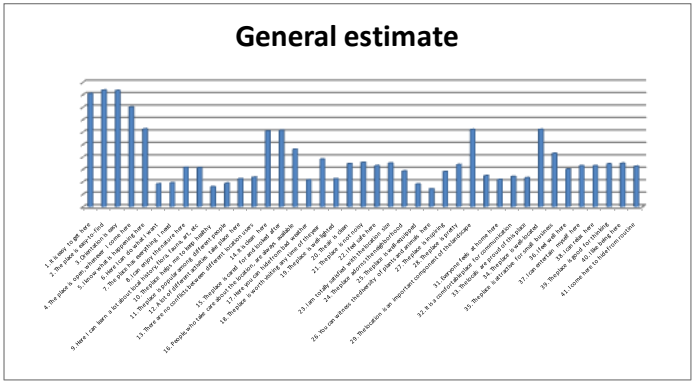


Fig. 3. Relative ratings of 41 bottom-level criteria

Aggregate relative ratings of third-level criteria are shown on Fig. 5, illustrating respondents' opinion about location's compliance with the parameters listed in the survey. As we can see from the diagram, the

weakest points of the location are that (according to the respondents) it does not meet the needs of the community and is rarely used.

Based on textual analysis of verbal reviews and "letters from the future", tag clouds were built. It is interesting to note that the drawbacks, mentioned by respondents in verbal reviews, confirm the ratings, shown on Fig. 5 (i.e., respondents' judgments are rather consistent): according to the respondents, low level of location usage and its inability to serve community interests, are the main flaws of the location.

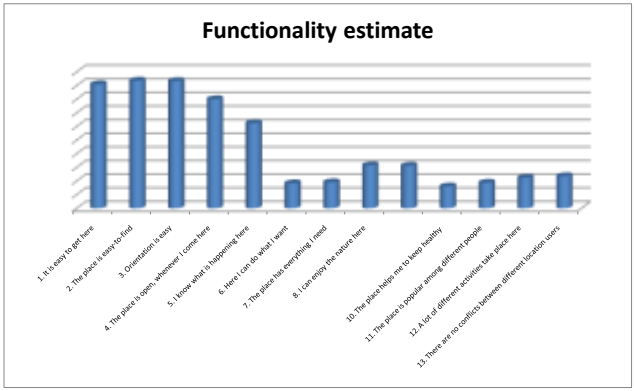


Fig. 4. Functionality of location: relative ratings of criteria

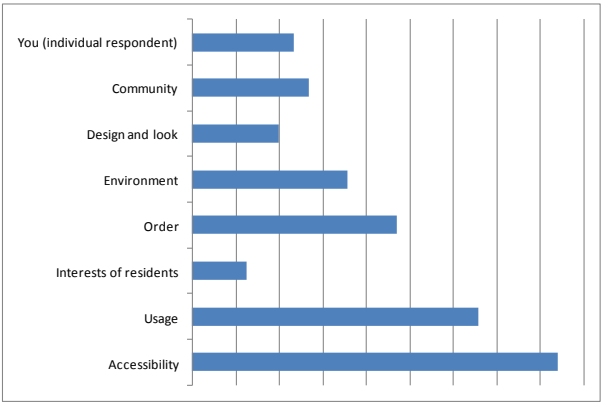


Fig. 5. Generalized ratings of third-level criteria

Aggregate survey results allow us to make several important conclusions.

1. Location has a great potential for further improvement and development. Community representatives consider themselves capable of active participation in transformation of the public space.

2. Museum space and adjacent territory are not the focal point of active community life. With the exception of museum employees, community residents rarely visit the location.

3. The main advantages of the public space are convenient downtown location, coziness, presence of greenery, open territory, and esthetic attractiveness of the historical museum building. The main drawbacks (and, consequently the main points for location development) include poor understanding of its designation by the community, lack of cultural events, activities, entertainment, management initiatives, creativity; alerting condition of trees on the territory of the museum.

4. While accessibility (convenient location), attractiveness for the community, and environmental potential are the strong points of the space, inclusiveness (consideration of interests of all potential user categories), full-fledged utilization of location's capacity, design/look, and infrastructure are the top-priority development aspects.

5. Transformation and sustainable development of the location should be based on the results of the conducted study, particularly, on aggregate judgments of the respondents from among community members. Location development activities and projects, implemented by local authorities, NGOs, activists, volunteers, based on public opinion, will be successful and get support from the community.

As of now, public organizations in cooperation with local activists and authorities have already accomplished several projects along the lines of the study results.

3 Peculiar features of the approach: place in decision science; advantages and disadvantages

As we can see, the described approach is a "hybrid" one in a way that it combines the elements of both decision theory and sociology (surveying, agreement scale usage).

The following features of the approach are common with the available decision support methods:

— usage of hierarchic problem decomposition (as in AHP [8] and CTDEA [11, 16]);
— heuristic transition from verbal judgments (like "agree/disagree") to numeric values (in the example – from "0" to "6"). Such a transition, in one form or the other, happens, virtually, in all multi-criteria alternative estimation methods that feature linear convolution (weighted summation) of ordinal or cardinal values (including Borda, Condorcet, AHP, TOPSIS, CTDEA, etc), because, as Litvak showed in [10], the necessary and sufficient condition of existence of an aggregate criterion (convolution across its sub-criteria) is expression of alternative estimates according to these sub-criteria in the ratio scale;
— aggregation (generalization, in our case through linear convolution) of data across multiple criteria, obtained from multiple respondents;
— verification of consistency of judgments (in our case – through informal analysis of verbal reviews).
Now let us list the main differences of the approach from the existing methods.
— the key task is only to describe the subject domain, i.e. to form a system of linked criteria (not to compare alternatives or projects according to these criteria, as it is done in AHP or CTDEA);
— the way of criterion formulation. "Atomic" bottom-level criteria, which do not have descendants in the hierarchy graph, are formulated as positive statements (and not definitions, as in traditional methods), with which a respondents can agree or disagree;
— as a result, instead of direct estimates or pair-wise comparisons, Likert's agreement scale is used;
— ease of problem decomposition: hierarchy graph is a "tree" [33, 34]; a network-type structure, i.e. a graph, in which any node can have more than one ancestor, can be too complex to be perceived by respondents;
— no need for coaching sessions with respondents (thanks to simplicity and transparency of the method).

So, we should stress once more that the key feature of the method is the combination of expert and sociological mindsets, which ensures, on the one hand, ease of use, and on the other – high efficiency of the method.
The results of the method's work are the ratings of activity scopes, providing the basis for further prioritization and, potentially, for allocation of limited resources [21].

The method's advantages are ease-of-use and understandability for a community member, efficiency and transparency, universality and flexibility (for each new subject domain a new unique hierarchy can be built), vividness of subject domain description process and representation of results.

The method's key disadvantages are the possibility of manipulations (experts can formulate criteria in some biased way, however, this is the issue of ethical principles of these experts and the DM), and of emergence of "lobbies" among community members (according to profession, age, mindset, gender, wealth, social status, etc).

A separate problem concerns manual input and processing of data (when MS Excel is the only software tool used for data aggregation). Ideally, the process of formal description of subject domain should be almost fully automated. Certain steps of the above-listed algorithm are already automated within the existing DSS. Particularly, "Consensus-2" DSS [42] includes means for registration of experts who are inputting the hierarchy, and for input of the hierarchy itself. "Solon" DSS [11, 41] includes tools for hierarchy input, as well as for calculation of relative impacts of criteria. Thus, the functions, delegated to experts, are already automated, while functions, delegated to respondents and to the research organizer (knowledge engineer, who has to aggregate the data and obtain recommendations for the DM using the DSS tools) still require automation. In order to simplify the process of surveying and aggregation of survey data is would be reasonable to automate:

— completion of the surveys (for example, using tablets or similar gadgets);

— submission of survey data to DSS knowledge base in remote mode;

— calculation of criterion ratings using the mathematical tools of a DSS.

Recognition and frequency analysis of verbal review texts, as well as tag cloud building are a bit more difficult to automate, however improvement of this algorithm step is no less relevant than automation of other steps.

4 Conclusions

It has been shown that community-level problems represent an example of weakly structured subject domains. During their formal description we should consider both expert data and community members' opinion. In view of the need to consider public opinion during

community-level decision-making, it is unreasonable to apply existing decision support methods and technologies in their classical form.

An efficient, yet simple, method has been suggested for formal description and analysis of community-level problems, taking public opinion into consideration. The method allows a DM, a local authority, an NGO, community representatives, volunteers, activists, media, or any other interested parties to get a clear understanding of a specific subject domain, which will provide the basis for prioritizing of future steps.

Experimental results have been obtained, based on conducted research of quality of a specific public space. The research confirms both efficiency and ease-of-use of the suggested approach. Based on the research results, specific recommendations concerning improvement of the target public space (location) have been worked out.

The described method should be used for community-level decision-making – in villages, raions, towns, neighborhoods — in spheres, immediately concerning the respective community members. Particular decisions might concern such aspects as planning and improvement of road and transport networks, domestic waste disposal, water supply and disposal, planning and improvement of territories, reintegration of public spaces into active community life, neutralization of negative information impacts, etc.

The method is an efficient decision support tool, which should be used by local self-government bodies, civil organizations, volunteers, activists, and any other interested parties.

Further studies will be dedicated to search for new applications of the method, and to automation of particular steps of the described procedure of analysis of weakly structured subject domains.

References

1. Kadenko S.V. Prospects and Potential of Expert Decision-making Support Techniques Implementation in Information Security Area; in Selected Papers of the XVIIIII International Scientific and Practical Conference on Information Technologies and Security (ITS 2016) Kyiv 2016/ CEUR Workshop Proceedings, pp. 8-14 (2016).
2. Boyd S., Vandenberghe L. Convex Optimization. Cambridge University Press (2004).
3. Keeney R.L., Raiffa H. Decisions with Multiple Objectives: Preferences and Value Tradeoffs Cambridge University Press. (1993).

4. Kendall, M. A New Measure of Rank Correlation. Biometrika. 30 (1–2), pp. 81–89 (1938).
5. Arrow K. J. Social Choice and Individual Values, 2nd ed. New York: Wiley (1963).
6. John G. Kemeny. Mathematics without Numbers. *Daedalus*. Vol. 88, No. 4, pp. 577-591 (1959).
7. Fishburn P.C. Utility Theory for Decision Making. Wiley (1970).
8. Saaty, T.L. Fundamentals of Decision Making and Priority Theory with The Analytic Hierarchy Process. RWS Publications, Pittsburgh PA (1994).
9. Миркин Б.Г. Проблема группового выбора. М. Наука (1974).
10. Литвак Б. Г. Экспертная информация. Методы получения и анализа. М.: Радио и свіязь (1982).
11. Тоценко В.Г. Методы и системы поддержки принятия решений. Алгоритмический аспект. ИПРИ НАНУ. К.: Наукова думка (2002).
12. Гнатієнко Г. М., Снитюк В.Є. Експертні технології прийняття рішень. К.: ТОВ "Маклаут" (2008).
13. Hwang, C. L., Yoon K. Multiple attribute decision making: methods and applications : a state-of-the-art survey. Berlin; New York : Springer-Verlag (1981).
14. Roy, Bernard. Classement et choix en présence de points de vue multiples (la méthode ELECTRE). La Revue d'Informatique et de Recherche Opérationelle (RIRO). (8): pp. 57–75 (1968).
15. M.Z.Zgurovsky, Yu.P.Zaychenko. The Fundamentals of Computational Intelligence: System Approach. Studies in Computational Intelligence, 652. Springer. (2005).
16. Циганок В. В. Удосконалення методу цільового динамічного оцінювання альтернатив та особливості його застосування. Реєстрація, зберігання і оброб. даних. Т. 15, № 1. С. 90-99 (2013).
17. Tsyganok V., Kadenko S., Andriichuk O., Roik P. Combinatorial Method for Aggregation of Incomplete Group Judgments. IEEE First International Conference on System Analysis & Intelligent Computing (SAIC), DOI: 10.1109/SAIC.2018.8516768, pp. 25-30 (2018).
18. Tsyganok V. Investigation of the aggregation effectiveness of expert estimates obtained by the pairwise comparison method. Mathematical and Computer Modeling. 52(3-4). pp. 538–544 (2010).
19. Proceedings of the International symposium for the analytic hierarchy process (archive). http://www.isahp.org/proceedings/, last accessed 2018/11/15.
20. Терентьєв О. М., Просянкіна-Жарова Т.І., Савастьянов В.В. Використання засобів текстової аналітики як інструменту оптимізації підтримки прийняття рішень у задачах розробки планів соціально-економічного розвитку України. Реєстрація, зберігання і оброб. даних..Т. 18, № 3. сс. 75-86 (2016).

21. Tsyganok. V., Kadenko S., Andriichuk O., Roik P. Usage of multicriteria decision-making support arsenal for strategic planning in environmental protection sphere. Journal of Multi-criteria Decision Analysis. Vol. 24, Issue 5-6, pp. 227-238 (2017).
22. Чепков І.Б., Ланецкий Б.М., Леонтьєв О.Б., Лук'янчук В.В. Методичний підхід до обґрунтування раціонального співвідношення обсягів розробки, закупівлі та ремонту озброєння й військової техніки. Озброєння та військова техніка. No 3. С. 9–14 (2014).
23. Циганок В.В., Каденко С.В., Андрійчук О.В., Качанов П.Т., Роїк П.Д. Інструментарій підтримки прийняття рішень як засіб стратегічного планування. Озброєння та військова техніка. № 3(7). С. 59-66 (2015).
24. Горбулін В.П., Додонов О.Г., Ланде Д.В. Інформаційні операції та безпека суспільства: загрози, протидія, моделювання: монографія. К., Інтертехнологія (2009).
25. Додонов А.Г., Ландэ Д.В., Цыганок В.В., Андрейчук О.В., Каденко С.В., Грайворонская А.Н. Распознавание информационных операций. Киев. (2017).
26. Kadenko S.V. Defining Relative Weights of Data Sources during Aggregation of Pair-wise Comparisons. Selected Papers of the XVII International Scientific and Practical Conference on Information Technologies and Security (ITS 2017) Kyiv 2017. pp. 47-55 (2017).
27. Закон України Про добровільне об'єднання територіальних громад (Відомості Верховної Ради (ВВР), 2015, № 13, ст.91). http://zakon5.rada.gov.ua/laws/show/157-19, last accessed 2018/11/15.
28. Розпорядження Кабінету Міністрів України від 1 квітня 2014 р. № 333-р «Про схвалення Концепції реформування місцевого самоврядування та територіальної організації влади в Україні». http://zakon0.rada.gov.ua/laws/show/333-2014-%D1%80, last accessed 2018/11/15.
29. Pankratova, N.D. & Nedashkovskaya, N.I. Hybrid Method of Multicriteria Evaluation of Decision Alternatives. Cybern Syst Anal 50: 701. https://doi.org/10.1007/s10559-014-9660-2 (2014).
30. N.D.Pankratova & N.I.Nedashkovskaya. A decision support system for evaluation of decision alternatives on basis of a network criteria model. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON). DOI: 10.1109/UKRCON.2017.8100363 (2017).
31. Каденко С.В. Проблеми представлення експертних даних у системах підтримки прийняття рішень. Реєстрація, зберігання і обробка даних. Т. 18 № 3, С. 67-74 (2016).
32. Tsyganok, V.V., Kadenko, S.V., & Andriichuk, O.V. Using Different Pair-wise Comparison Scales for Developing Industrial Strategies. Int. J. Management and Decision Making. 14(3), pp 224-250 (2015).

33. Saaty, T.L. The Analytic Hierarchy Process. McGraw-Hill, New York. (1980).
34. Kadenko S.V. Determination of Parameters of Criteria of «Tree» Type Hierarchy on the Basis of Ordinal Estimates. Journal of Information and Automation Sciences. Vol. 40. i.8. P. 7–15 (2008).
35. Likert R. A Technique for the Measurement of Attitudes. *Archives of Psychology*. # 140. P. 1–55 (1932).
36. Olenko A., Tsyganok V. Double Entropy Inter-Rater Agreement Indices. Applied Psychological Measurement 40(1). pp. 37–55 (2016).
37. SpaceShaper User's Guide. https://www.designcouncil.org.uk/resources/guide/ spaceshaper-users-guide, last accessed 2018/11/15.
38. St. James Park, Southampton. CABE Spaceshaper Workshop Facilitators Report. http://www.westleydesign.co.uk/what-we-do/Downloads/WestleyDesign-StJamesSpaceshaper.pdf, last accessed 2018/11/15.
39. Camberwell Town Centre Spaceshaper Consultation Report. https://www.southwark.gov.uk/assets/attach/4070/Camberwell_Town_Centre__Spaceshaper_Consultation_Report_December_2011.pdf, last accessed 2018/11/15.
40. Carawatha Park Spaceshaper Workshops. http://www.melvillecity.com.au/static/ attachments/2013/April/3383_SpaceShaper_Report.pdf, last accessed 2018/11/15.
41. Тоценко В.Г., Качанов П.Т., Циганок В.В. Комп'ютерна програма «Система підтримки прийняття рішень «Солон-3» (СППР «Солон-3»). Свідоцтво про реєстрацію авторського права на твір № 8669 від 31.10.2003 (2003).
42. Циганок В.В., Роїк П.Д., Андрійчук О.В., Каденко С.В. Комп'ютерна програма «Система розподіленого збору та обробки інформації для систем підтримки прийняття рішень «Консенсус-2» (СРЗОІ «Консенсус-3»). Свідоцтво про реєстрацію авторського права на твір № 75023 від 27.11.2017 (2017).

SECURITY ESTIMATION OF THE SIMULATION POLYGON FOR THE PROTECTION OF CRITICAL INFORMATION RESOURCES

Bogdan Y. Korniyenko[1], Liliya P. Galata[2], Lesya R. Ladieva[3]
[1] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute",
Kyiv, Ukraine, e-mail: bogdanko@i.ua
[2] Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, e-mail: galataliliya@gmail.com
[3] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute",
Kyiv, Ukraine, e-mail: lrynus@yahoo.com

In the article the question of Security estimation of information system protection through risk analysis is considered. An analysis of information risks is conducted for testing information security system, which allows to identify threats to information security. At present, different methods of analyzing information risks exist and are used, the main difference of which is in the scale of risk assessment: quantitative or qualitative. Based on analyzed existing methods of testing and assessing the vulnerabilities of the automated system, their advantages and disadvantages, for the possibility of further comparing the spent resources and information system security, a conclusion is made for the definition of an optimal method of testing the information security system method in the context of a constructed simulation polygon for the protection of critical information resources. The simulation polygon for the protection of critical information resources was developed and implemented based on the GNS3 application software. It is also concluded that the assessment of network security with mixed (complex) methods is not feasible. The optimal iRisk methodology for testing the information security system based on the simulation polygon for protection of critical information resources has been identified, among the considered methods for testing and analysis of automated system risks. The quantitative method iRisk is considered for Security estimation of information system protection. The general risk assessment iRisk is calculated considering the following parameters: Vulnerability Assessment, Threat Assessment, assessment of security tools. The methodology contains the general CVSS v3 vulnerability assessment system, which allows you to use constantly relevant coefficients to calculate vulnerabilities, and also have a list of all the major vulnerabilities that are associated with all modern software products that can be used in the automated system. The known vulnerabilities of used software and hardware are considered and the

stability of the built simulation polygon for the protection of critical information resources to specific threats is calculated by iRisk method.

Keywords: *Simulation Polygon, Critical Information Resources, Security, Vulnerability, Threat, Control.*

Introduction

Periodic analysis of information risks is conducted for the research of information security system, it allows to identify threats to information security and in turn use and implement appropriate measures for their neutralization [1].

Based on the research and development of the simulation polygon for the protection of critical information resources by GNS3 application software, we can conclude that testing and evaluation of the constructed a secure network should be considered in the context of testing performance, impacting settings on the automated system security level, and in the context of used information protection tools [2]. This is due to the fact that in this case the emphasis is on the technical part, practically not considering organizational measures related to information security in the AS. Given that the emphasis is on hardware, software and network level of information protection, so network security evaluation by mixed (complex) methods is not appropriate.

Based on the fact that quantitative methods in conducting a risk analysis at software and technical protection level and if not consider organizational and technical component, are more effective, it should choose a quantitative evaluation method of protection [3, 4].

Among the main quantitative methods for analyzing information risks RiskWatch, Digital Security, ISRAM and iRisk, the iRisk method is more acceptable. The reason for this is, first of all, that this technique is free, informative enough, includes another CVSS v3 vulnerability assessment method, which is actively supported by the National Institute of Standards and Technology, and contains up-to-date information about the critical vulnerabilities of software and hardware, which in turn allows for an effective assessment of the level of network security.

The task that needs to be solved is to research of the simulation polygon for the protection of critical information resources by iRisk method for effectively assess the level of network security, considering the fact that the emphasis is on the hardware-software and network levels of information security.

1 iRisk method

The iRisk method is formally one of the simplest estimates of information security quantitative risks for automated system. In general, it is calculated by the following equation:

$$iRisk = (Vulnerability \cdot Threat) - Controls \quad (1)$$

where *Vulnerability* - vulnerability assessment, *Threat* - threat assessment, *Control* - assessment of security tools. This technique uses a different Common Vulnerability Scoring System v3.0 (CVSS V3) methodology for vulnerability assessment.

When assessing the threat, the probability of realization of the threat and the degree of its influence are being assessed. The degree of impact of the threat is estimated through the indicators of losses. To assess the probability of implementing a specific threat, there are two indicators: ARO is the expected number of threats during the year, and the level of knowledge and the offender's access level in the AS.

Formally, the calculation is not a complicated equation, but this methodology contains a general CVSS vulnerability assessment system, which is supported by market leaders in the field of information security in practice, that allows you to use constantly relevant coefficients for calculating vulnerabilities, and also have a list of all the major vulnerabilities associated with all modern software products that can be used in an automated system [5].

1.1 Vulnerability

First of all, we have calculated Vulnerability, by using the standard CVSS v3. The calculation takes place according to the scheme presented in Fig. 1. During the calculation, a large number of coefficients are used, so for convenience we will use the software of the National Institute of Standards and Technologies, then correct parameters setting will allow to get the result of calculations in the form of a scale from 1 to 10, where 1 it's a low level (no vulnerability), and the value 10 it's the critical vulnerability that needs to be eliminated. The standard includes three groups of metrics required for calculation: base, temporal and environmental.

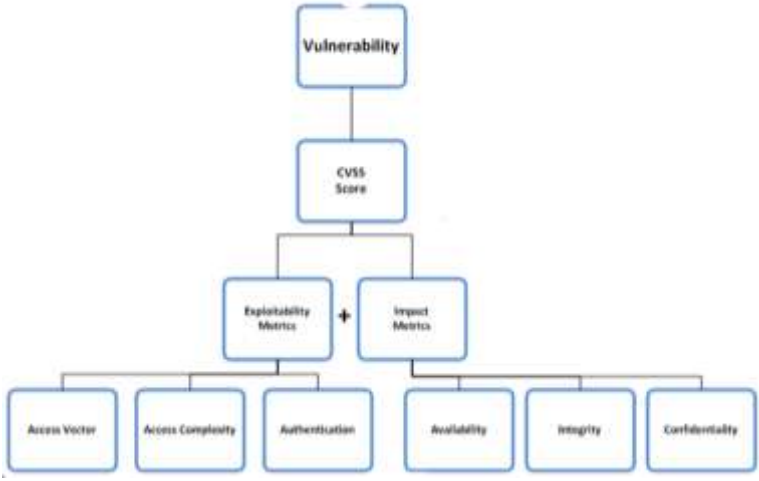


Fig. 1. General scheme of vulnerability calculation in CVSS v3

The value of the metric is accepted as a pair of vector (specific values of individual indicators) and a numerical value, which is calculated basing on all indicators and using the equation defined in the standard. Fig. 2 shows all the necessary parameters for calculating the environmental metric of the polygon for the protection of critical information resources.

Fig. 2. The environmental metric of the polygon for the protection of critical information resources

1.2 Threat Assessment

According to this standard, the threat is explained as a negative event that may result of the vulnerability benefits. In order to make the equation as simple as possible, the iRisk method focuses on two main components: *impact* and *likelihood*. Fig. 3 is presented the scheme of threats estimation in iRisk method.

Fig. 3. Scheme for threats estimating in the iRisk method

Impact is the amount of damage that this incident will bring to the organization. Within the iRisk SecureState equation, today the following criteria are used to determine the impact. By default, the following values are assigned, but they can be changed according to the needs of the evaluated object:

- financial (25) - whether threats destroy the organization financial flows;
- strategic (15) – whether threats lead to long-term strategic losses;
- operational (25) – whether threats influence on the work continuity;
- law compliance (25) - whether threats affect the ability to keep to the standards;
- reputation (10) - whether threats affect the relationship with customers.

Likelihood is another major component of the threat. The iRisk method uses two factors to estimate the probability: the annual expected number of threat implementations and the attacker's level of knowledge and access (correlation table between the level of knowledge/access and the annual number of threat implementations ARO (annualized rate of occurrence) [6]).

The threat is calculated by the Eq. (2), where Likelihood (correlation from table ARO [6]). If the threat is on a scale from 100 to 50 - the level of risk is high, from 50 to 10 – medium**,** from 1 to 10 - low.

$$Threat = Impact \cdot LikeLihood \qquad (2)$$

1.3 Control (assessment of security tools)

Based on the definition of the ISACA organization, preventive, detection, correction or deterrence means for security may be used in iRisk. The structure of the Control parameter (assessment of security tools) is presented on Fig. 4.

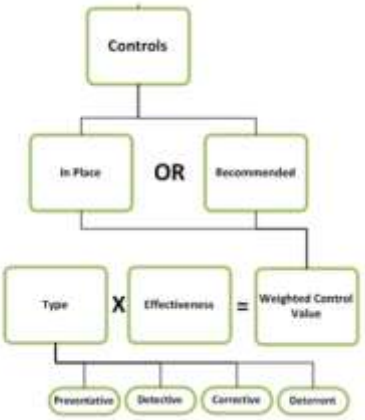


Fig. 4. Structure of the Control parameter of the iRisk method

According to the standard, the tools have the following ratings: preventive - 5, detection - 4, correction - 3, deterrence -3.

The next step is to define the *Controls* (efficiency), it has a five-point scale by the standard: 5 - if the information security tools in the network significantly exceed the goal, 4 - exceed the goal, 3 - the implementation corresponds to the goal, 2 – the implementation is not fully satisfying its goal, 1 - slightly up to its goal.

Adding indicators by CVSS we will get the following values:

- optimized (801 - 1000) - the tool can't be developed or implemented better;
- managed (601 - 800) - the tool continues to improve;
- defined (401 - 600) - the security tools are clearly defined and reduce the risk to medium;
- initial / Ad-Hoc (1 - 200) – the tool provides only some protection value.

Thus, the three main components, which appears in the method iRisk, balance each other. The highest possible score for the threat is 100, which is multiplied by the maximum vulnerability (10). That is 1000 points potential, which is compensated by the potentially perfectly implemented protection, at the end will leave zero risk. In practice, this is almost not achievable and, in any case, left a part of the residual risk. That is, the risk varies in values from 0 to 1000, in this case the smaller value means the more secure automated system.

1.4 Software and hardware vulnerabilities

The designed simulation cybersecurity polygon hasn't so many vulnerabilities due to the high-quality equipment, the access control that divides the network into the demilitarized zone, the internal and external network, and the network settings, that limit access to the network from the outside, limit number of half-connections, which reduces the effectiveness of DdoS attacks, network scan, etc. [2]. And still, the vulnerabilities remain on the software and hardware level. Next, we will look at some of them, the calculation of the security of the polygon for the protection of critical information resources will be done using iRisk [7-10].

1.5 Cisco IOS Arbitrary Command Execution Vulnerability (CVE-2012-0384)

The vulnerability occurs due to error in HTTP/HTTPS authorization that allows an authenticated user to execute any Cisco IOS software commands configured for user privilege levels.

We will calculate the base metric for Vulnerability calculation, and for more correctness, according to the security of the cybersecurity polygon

we will calculate the temporal and environmental metric, as described above [11-14].

Base Score Metrics {Attack Complexity = Low; Privileges Required = Low; User Interaction = None; Scope= Unchanged; Confidentiality Impact = High; Integrity Impact = High; Availability Impact = High}

Temporal Score Metrics Score Metrics {Exploitability = Functional exploit exist}

Environmental Score Metrics {Base Modifiers {Attack Vector = Local; Attack Complexity = Low; Privileges Required = Low; User Interaction = None} {Scope = Unchanged} {Impact Metrics {Confidentiality Impact = Low; Integrity Impact = Low; Availability Impact = High}} {Impact Subscore Modifiers {Confidentiality Requirement = Low; Integrity Requirement = Low; Availability Requirement = Low}}}

The resulting calculation of the base level Vulnerability assessment equal 7.8 out of 10, which is shown on Fig. 5.

Considering that the threat should be realized from inside and first of all is oriented to a normal user without administrator rights and the expected number of threats is estimated as high, then from the ARO table [6] we choose the correlation value Impact = 0.9. So, according to the Eq. (2): Threat = 0.9 · 100 = 90.

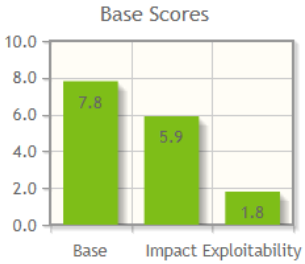


Fig. 5. The Base CVE-2012-0384 vulnerability metric for the cybersecurity polygon

As described above, the value Controls is estimated at 650, which will mean - the tool continues to improve.

That is, the value iRisk = (7.8 · 90) - 650 = 50 for Cisco IOS Arbitrary Command Execution Vulnerability (CVE-2012-0384).

1.6 Cisco Access Control Bypass Vulnerability (CVE-2012-1342)

The vulnerability of Cisco routers allows remote attacks to bypass the Access Control List (ACL) and send network traffic that should be rejected. Implementation of vulnerability leads to a violation of the automated system integrity.

In the same way as for the CVE-2012-0384 vulnerability, we will calculate the iRisk value.

Base Score Metrics {Attack Vector = Network; Attack Complexity = Low; Privileges Required = None; User Interaction = None; Scope= Changed; Confidentiality Impact = None; Integrity Impact = Low; Availability Impact = Impact None}

The value Vulnerability = 5.8, by the CVSS v3.0 calculator (Fig. 6).

The calculation of the value Threat = 1.4 · 0.72 · 100 = 108, so the value iRisk = (5.8 · 108) - 610 = 16.4, which means that the vulnerability will be approximately equal to zero, that is we can conclude that this vulnerability can be exploited by an attacker with little probability.

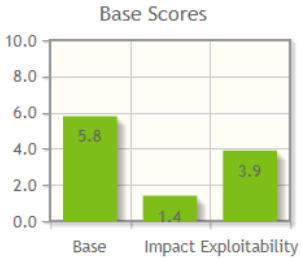


Fig. 6. The Base CVE-2012-1342 vulnerability metric for the cybersecurity polygon

1.7 EternalBlue vulnerability (CVE-2017-0144)

This vulnerability uses the vulnerability in the implementation of the Server Message Block v1 protocol (SMB). An attacker, having formed and transmitted to a remote host a specially prepared package, is able to get remote access to the system and run any code.

Calculate the iRisk value for CVE-2017-0144 EternalBlue vulnerability.

The base EternalBlue vulnerability metric will have the following parameters. The result is shown in Fig. 7

Base Score Metrics {Attack Vector = Network; Attack Complexity = High; Privileges Required = None; User Interaction = None; Scope= Unhanged; Confidentiality Impact = High; Integrity Impact = High; Availability Impact = High}

Since the attack is conducted from the outside and its' probability is very high, the attacker should be an hacking expert, according to the iRisk method in this case, the value Impact = 100, and the value Likelihood = 0.7 and the value Threat =70,

So, you can calculate the iRisk value for CVE-2017-0144, without the security patch from March 14, 2017: iRisk = (8.1 × 70) - 0 = 567.

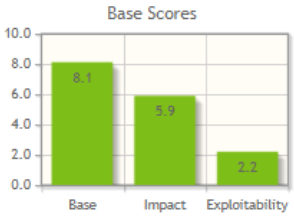


Fig. 7. The Base CVE-2017-0144 EternalBlue vulnerability metric for the cybersecurity polygon

1.8 Meltdown vulnerability (CVE-2017-5754)

Vulnerability exploits the effect of out-of-order execution in modern processors. Attack doesn't depend on the operating system and doesn't exploit software vulnerabilities. Meltdown actually breaks down the entire security system based on the isolation of the address area, including the virtual one. Meltdown allows you to read part of the memory of other processes and virtual machines. The KAISER patch excludes this vulnerability, but reduces CPU performance.

Calculate the iRisk value for a cybersecurity polygon, without KAISER patch.

Calculate the base metric for Meltdown vulnerability (CVE-2017-5754), the result is shown in Fig. 8.

Base Score Metrics {Attack Vector = Local; Attack Complexity = High; Privileges Required = Low; User Interaction = None; Scope= Changed; Confidentiality Impact = High; Integrity Impact = None; Availability Impact = Impact None}

Considering that the attacker can act both from the outside and inside and the attack can be executed frequently, and the attacker can have just an advanced level of skills and the attack code is shown in large numbers

of articles, all of this will give a correlation value of Impact = 0.9, and the value of Threat will be equal to 100 · 0.9 = 90.

The resulting value of iRisk for Meltdown (CVE-2017-5754) will be equal to iRisk = (5.6 · 90) -0 = 504, because without the KAISER patch this Vulnerability doesn't show itself, and is included in the architecture of most modern processors.

Fig. 8. The Base CVE-2017-5754 Meltdown vulnerability metric for the cybersecurity polygon

1.9 SPECTRE vulnerability (CVE-2017-5753, CVE-2017-5715)

This vulnerability is assigned two identifiers CVE-2017-5753, CVE-2017-5715. By its nature, it is similar to Meltdown, but with some differences, in particular, by during a speculative code execution, the processor can execute instructions that it would not perform under strictly consistent (non-speculative) calculations, and although in the future the result of their performance is discarded, its imprint remains in the processor cache and can be used.

The Specter vulnerability is not easy to implement - however, it can be implemented, under the condition of attack on a specific software, known to the attacker and, if possible, available in an open source code in the same version and on the same system, which provides an attack.

Another way for Specter implementing is to "predict branching" - the processor has a similar transition prediction block, it predicts the transition address for the next instruction of the indirect transition (Meltdown, but here they play a different role).

For simplicity, this unit does not broadcast between virtual and real addresses, which means it can be trained in the address space of the attacker on certain actions.

After some time, the real transition address will be deducted, the processor identifies the error and rejects the results of the speculative

execution, however, as in all other instances of the use of Meltdown and Specter, most performance results remain in the cache.

Calculate the iRisk value for the Specter vulnerability. The base metric in both versions of the vulnerabilities implementation is the same, the results of the calculation are presented in Fig. 9.

Base Score Metrics {Attack Vector = Local; Attack Complexity = High; Privileges Required = Low; User Interaction = None; Scope= Changed; Confidentiality Impact = High; Integrity Impact = None; Availability Impact = Impact None}

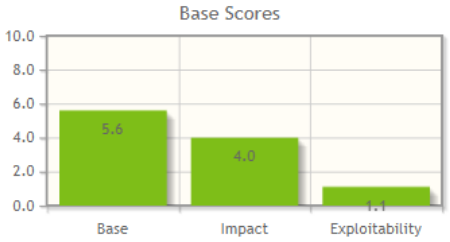


Fig. 9. The Base Spectre CVE-2017-5753 i CVE-2017-5715 vulnerability metric for the cybersecurity polygon

In both cases with Spectre, we are concerned with the fact that the processor learns fast to execute one process by using as an example another process, thereby actually allowing the second process to control the progress of the first one. There are no universal patchs to fix Specter, and ways of protection from CVE-2017-5715 are the permanently clearing the cache and cleaning the code from the core.

Calculate the iRisk value for CVE-2017-5715, given the complexity of the exact implementation and the impact only on the information confidentiality. So the value of Impact = 50 (including financial, reputational and strategic impact). Given that the vulnerability will be try to use mainly from the outside and the attacker must have advanced technical skills, the correlation value Likelihood = 0.64. These parameters are typical for both CVE-2017-5753 and CVE-2017-5715.

However, the Controls parameters in this case need to be evaluated in different ways. There are patches for CVE-2017-5715 vulnerability, which partially solve this problem only in some cases, so value Controls can be considered Initial/Ad-Hoc = 100, but it's provides only some protection value. As to CVE-2017-5753 vulnerability, value Controls can be considered as 0, as this problem is not resolved at this time.

So, for CVE-2017-5715 iRisk = (5.6 · 50 · 0.64) - 100 = 79.2.

For CVE-2017-5753 iRisk = (5.6 · 50 · 0.64) - 0 = 179.2

2 Conclusions

The iRisk method was chosen for the research, first of all because this technique is free, enough informative, includes another CVSS v3 vulnerability assessment method, which is actively supported by the National Institute of Standards and Technology. Automated system has been tested for the following vulnerabilities: Cisco IOS Arbitrary Command Execution Vulnerability (CVE-2012-0384), Cisco Access Control Bypass Vulnerability (CVE-2012-1342), EternalBlue (CVE-2017-0144), Meltdown (CVE-2017-5754), Specter (CVE-2017-5753) (CVE-2017-5715). Conclusions have been shown about the stability of the designed network to specific threats by the iRisk method. It uses the values from 0 to 1000 scope, where 0 corresponds to automated system, in which it is possible to neglect this vulnerability, whereas at the maximum value, if it exceeds 100, it is necessary to solve this vulnerability. The results of calculations are given in Table 1.

Table 1. Table of iRisk values for a builted cybersecurity polygon

Vulnerability	***Value iRisk***
Cisco IOS Arbitrary Command Execution Vulnerability (CVE-2012-0384)	50
Cisco Access Control Bypass Vulnerability (CVE-2012-1342)	16.4
EternalBlue (CVE-2017-0144)	567
Meltdown (CVE-2017-5754)	504
Spectre (CVE-2017-5715)	79.2
Spectre (CVE-2017-5753)	179.2

The higher the value iRisk the vulnerability is the more critical and has a higher priority for automated system protection.

References

1. Klaus Wehrle, James Gross. Modeling and Tools for Network Simulation. Hardcover: 256 p. (2010).
2. Korniyenko, B. Model of Open Systems Interconnection terms of information securitv. Science intensive technology, № 3 (15), pp. 83 – 89., doi.org/10.18372/2310-5461.15.5120 (ukr) (2012).

3. Korniyenko, B., Yudin, O., Novizki,j E. Open systems interconnection model investigation from the viewpoint of information security. The Advanced Science Journal, issue 8, pp. 53 – 56. (2013).
4. Korniyenko, B., Yudin, O. Galata, L. Research of the Simulation Polygon for the Protection of Critical Information Resources. CEUR Workshop Proceedings, Information Technologies and Security, Selected Papers of the XVII International Scientific and Practical Conference on Information Technologies and Security (ITS 2017), Kyiv, Ukraine, November 30, 2017, Vol-2067, - P.23-31, urn:nbn:de:0074-2067-8 (2017).
5. Chris Clymer, Ken Stasiak, Matt Neely, Stephen Marchewitz. IRisk Equatuion Available via https://securestate.en/iRisk-Equation-Whitepaper.pdf
6. Common Vulnerability Scoring System v3.0: User Guide. Available via https://www.first.org/cvss/user-guide
7. Korniyenko, B., Yudin, O. Implementation of information security a model of open systems interconnection. Abstracts of the VI International Scientific Conference "Computer systems and network technologies» (CSNT-2013), p. 73. (2013).
8. Korniyenko, B. Information security and computer network technologies: monograph. ISBN 978-3-330-02028-3, LAMBERT Academic Publishing, Saarbrucken, Deutschland, 102 p. (2016).
9. Korniyenko, B., Galata, L., Kozuberda, O. Modeling of security and risk assessment in information and communication system. Sciences of Europe, V. 2., No 2 (2), pp. 61 -63. (2016).
10. Korniyenko, B. The classification of information technologies and control systems. International scientific journal, № 2, pp. 78 - 81. (2016).
11. Korniyenko, B., Yudin, O. Galata, L. Risk estimation of information system. Wschodnioeuropejskie Czasopismo Naukowe, № 5, pp. 35 - 40. (2016).
12. Korniyenko, B., Galata, L., Udowenko, B. Simulation of information security of computer networks. Intellectual decision making systems and computing intelligence problems (ISDMCI'2016): Collection of scientific papers of the international scientific conference, Kherson, Ukraine, pp. 77 - 79. (2016).
13. Korniyenko, B. Cyber security - operating systems and protocols. ISBN 978-3-330-08397-4, LAMBERT Academic Publishing, Saarbrucken, Deutschland, 122 p. (2017).
14. Korniyenko, B., Galata, L. Design and research of mathematical model for information security system in computer network. Science intensive technology, № 2 (34), pp. 114 - 118. (2017).

PROTECTION OF INFORMATION NETWORKS BASED ON LORA TECHNOLOGY

Dmytro Kucherov[1], Andrii Berezkin[2]
[1] National Aviation University, Kiev, 03058, Ukraine
[2] Pukhov Institute for Modeling in Energy Engineering, 03164, Kiev, Ukraine
d_kucherov@ukr.net

The paper deals with modern technology for transmitting short messages over long distances named LoRa, where the transmitted signal uses linear frequency modulation (chirp). The object of the study to define lack of transmitters that it has a design on LoRa technology for assessment their applicable in condition urban city where there are a lot of radiation sources. The goal of the work is the creation of a method of assessing the act the interference conditions that based on measurement bit error rate and signal-noise ratio and via on which to get individual host vulnerability levels. The processing of these signals is carried out by means of a time-frequency transformation. The chirp signal is characterized by 4 parameters: frequencies, time, modulation rate and amplitude. By analogy with the wavelet transform, the processing of chirp signals involves a chirplet decomposition. Since the chirp signals are strongly influenced by mutual interference due to multipath, the article studies the effectiveness of LoRa technology in conditions of mutual interference of radiation sources. The developed method utilized chirplet decomposition and retrieve symbols of a message in the dictionary. The conducted experiments have confirmed the proposed software operability and allow recommending it for use in practice for solving the problems receiving signal. The prospects for further research may include the creation of parallel methods for calculation of the set of proposed indicators, the improvement of software, as well as an experimental study of proposed indicators in real conditions.

Keywords: *Chirp, Time-Frequency Transform, Chirplet Decomposition, LoRa Technology, Interference.*

Introduction

Currently, LoRa technology is widely used to create a number of devices for the Internet of things, data collection and transmission systems, and also portable devices, through which short messages can be transmitted over long distances (according to some data up to 20 km). An additional advantage of the technology is the conservation of energy resources of the user devices (galvanic cells). This technology uses chirp pulses, the message symbols in which are coded by 4 parameters such as

amplitude, center frequency, deviation range and the center of the pulse. Each symbol is assigned a chirp, characterized by these 4 parameters. A complete set of symbols forms a dictionary of chirplets, named by analogy with wavelets. The analysis of the message consists of selecting symbols from the dictionary and finding the best matching of the chirplets to the symbols of the received message. This analysis, called a chirplet decomposition, is carried out in the time-frequency domain. In spite of the fact that chirp signals use the spreading of the spectrum the jamming in the receiving channel cause some difficulties. Processing of chirp signals is complicated by mutual interference from similar sources, as a result of which the received signal has distortions. The main goal of the paper is to research the effectiveness of the chirplet decomposition under interference conditions.

1 Review of the Literature

Previous work [1] addressed general issues of network congestion assessment, without reference to vulnerability. However, solution providers, operators, and researchers show a natural interest in the latest network technology LoRa. The most detailed analysis of this technology is presented in [2]. This paper deals with some open questions related to LoRa research and development. The innovative mathematical model of the network LoRaWAN is presented in [3]. This model provides a determination of the network capacity and reliability of information transmission. Mathematical simulation of the radio channel for the LoRaWAN transceiver in various operating states for different environments covering the urban, suburban and rural areas is given in [4]. The study in [4] has shown that the best suitable model for all registered levels of the received measurement signal is described by the Nakagami distribution and, in general, LoRa is a reliable portable wireless technology. Asynchronous protocol LoRaWAN by type ALOHA for access to the channel without the limitation of the working cycle is presented in [6].

However, the processing of LoRa signals is currently not fully researched. The quality of signal processing is based on the research of their types and the methods of protection against interference that are used. It should be noted that the use of frequency analysis alone, in conjunction with classical digital processing, as it was used in [6], is not suitable since LoRa uses spread spectrum technology. Measurement of only the center frequency of the signal is not sufficient to decipher the

complete message. General information on the modulation used in LoRa technology is given in [7]. The main technology is the use of signals with linear frequency modulation. Processing chirp signal based on decomposition by Gaussian chirplets was an active area of research in signal processing in the 90s of the last century [8, 9].

The approach to chirplet-decomposition of the received signal is presented in [10]. However, the expansion of the adopted chirp on the basis of Gaussian functions turned out to be unsuitable, since Gaussian chirplets do not form an orthogonal basis. A promising solution was the scheme of decomposition of signals based on matching. An algorithm for searching of optimal Gaussian chirplets using a crude dictionary is presented in [11]. A similar algorithm for estimating the characteristics of visually evoked potentials (VEP) based on the Chirplet representation is applied in [12]. In [13, 14] the structure of the Chirp-Binary Orthogonal Keying (BOK) system is studied on a background of white Gaussian noise and a frequency filter for eliminating slit-like interference in direct-spectrum communications (DSC).

The main result of [13] is the estimation of the number of erroneous bits (BER) in the Chirp-BOK system. The protection of the receiving channel based on the decomposition of chirplet is presented in [14]. This paper also shows some chirplet-decomposition possibilities under interference conditions. The obtained results can be used in many areas including systems of communications of unmanned aerial vehicles for monitoring objects of different type [15].

2 Problem Statement

Chirp is a signal that has the form

$$s(t) = \begin{cases} a(t)\cos(2\pi ft + ct^2) \text{ if } |t| \le \tau_0 / 2; \\ 0, \text{ otherwise,} \end{cases} \quad (1)$$

where $a(t)$ is the law of amplitude variation (envelope); f is the central frequency; t is time, and c is the phase modulation coefficient. If $c > 0$, the frequency increases, if $c<0$, the frequency decreases, $c = 0$ corresponds to a harmonic signal that is not modulated in frequency; τ_0 is the pulse duration.

The received signal can be presented as an additive mixture of a useful signal $s(t)$, a white noise with zero mean $n(t)$ and a signal of re-reflections $w(t)$

$$y(t) = s(t) + n(t) + w(t) \qquad (2)$$

The main indicators of information network security are topological characteristics, one of which is a host's vulnerability. Host vulnerability is determined based on known vulnerabilities and the main type of vulnerability for LoRa system is the interference for receiving set. To assess the quality of reception, we use the signal-to-noise ratio and bit error rate and estimate the effectiveness of the LoRa system under consideration by measuring the signal-to-noise ratio in a densely populated urban area.

3 Elements of Protection

LoRa technology has a few elements of protection. First of them is a signal spectrum that it is completely determined by the phase modulation component and represents the Fourier transform of the signal *s* (*t*), i.e.

$$S(f) = \int_{-\infty}^{\infty} s(t) e^{-j2\pi ft} dt \ . \qquad (3)$$

For the signal *s*(*t*), representing a rectangular pulse of unit amplitude, i.e. $a(t) = 1$ and duration τ_0, expression (2) can be written in the form

$$S(f) = \int_{-\tau_0/2}^{\tau_0/2} e^{jct^2} e^{-j2\pi ft} dt \qquad (4)$$

The spectrum of the signal τ_0 =10 μs and bandwidth 200 MHz of the form (1) is shown in Fig. 1.

It should be noted that the width of the spectrum is an indirect indicator of security since high-density interference is difficult to create in a wide frequency range.

The next element of protection is encoding a chirp signal. In decoding used to use chirplet decomposition.

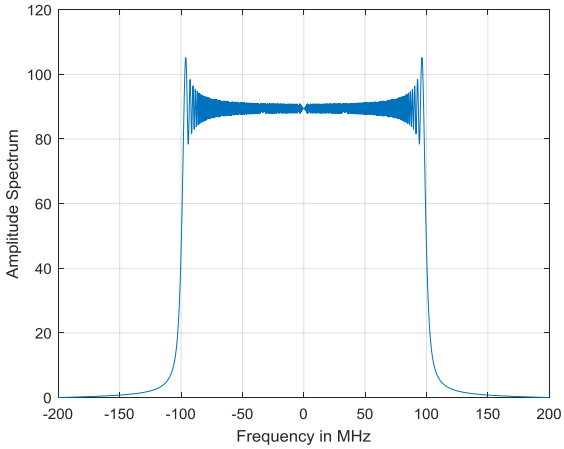


Fig. 1. Typical spectrum for chirp signal.

Chirplet is a Gaussian function of the form

$$g_c(t) = \frac{1}{\sqrt{\sqrt{\pi}\sigma}} \exp\left(-\frac{(t-t_c)^2}{2\sigma}\right) \exp(j2\pi f_c(t-t_c) + \mu_c(t-t_c)^2), \quad (5)$$

where t_c, f_c are parameters of the time and frequency of the function; σ is the variance, which determines the duration of the chirp function; and μc is the modulation rate. The Gaussian chirplet is a fundamental function. Therefore, it is desirable to represent the receiving signal during its processing as a weighted sum of Gaussian chirplet. The form of the chirplet signal is shown in Fig. 2.

For the parameters $t_c = 0$, $\sigma = 1$, $\mu_c = 0$, $f_c = 0$, the function $g_c(t)$ takes the form

$$g_c(t) = \frac{1}{\sqrt[4]{\pi}} \exp\left(-\frac{t^2}{2}\right) \quad (6)$$

Expression (6) is called the base function of transformations. Modification (6) can be used for the identification of parameters of the Gaussian function for its representation by a harmonic oscillation with frequency modulation of a given kind.

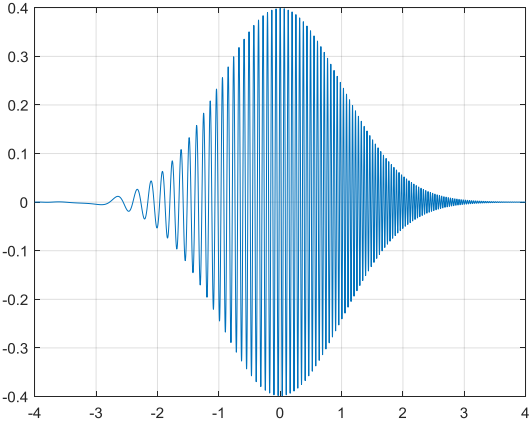


Fig. 2. The form of a chirplet

A set of chirplets can be used for the representation of chirp particles. For this purpose, in [8] it was suggested to use time convolution, frequency multiplication together with time and frequency shifts. Unfortunately, this approach to chirplet transformation does not give a positive result, because these transformations are interdependent, therefore such a chirplet cannot be chosen as a basis for orthogonal functions [10].

Recently, a direction has been developed, related to the development of the Fourier transform, in particular, a fractional Fourier transform, which measures the angular distribution of the signal energy in the time-frequency plane. This operator, given by the Wigner distribution for the signal $s(t)$

$$W(t,\omega)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} s\left(t-\frac{\tau}{2}\right)s^{*}\left(t+\frac{\tau}{2}\right)e^{-j\omega\tau}\,d\tau \qquad (7)$$

rotates the signal in the time-frequency plane. In formula (7) asterisk for signal $s^*(t)$ means the complex conjugate signal $s(t)$, ω is the angular frequency equal to $2\pi f$. Rotation represents a special combination of chirp convolution and of multiplication chirp as a result of an orthogonal transformation of time-frequency coordinates. This property is achieved by multiplying on the scale factor, using rotation, and by the time and

frequency shifts that form the four time-frequency atom parameters. However, it is possible to use the chirplet decomposition to represent the complex signal in a compact form. The appropriate time-frequency transform is presented in Fig. 3.

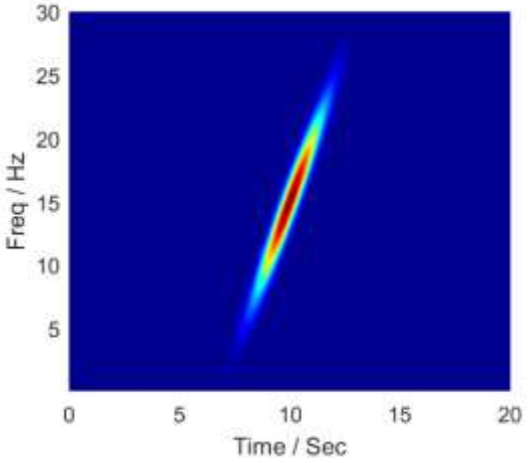


Fig. 3. Time-frequency transform of a chirplet for signal a = 1 V, f = 15 Hz, c = 15, τ_0 = 20 sec

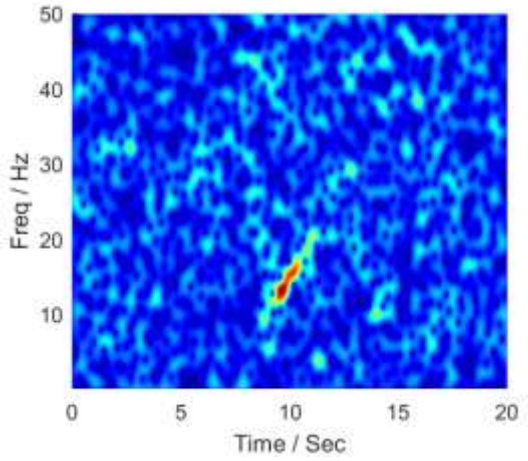


Fig. 4. Time-frequency representation for chirp signal in noise, SNR=2

An effective measure of the quality of the received data packet in an interfering noise environment is the probability of a transmission error of the data packet p_p, which can be expressed by the relation

$$p_p = 1-(1-p_e)^N, \qquad (8)$$

where p_e is the bit error probability of the information bit or bit error rate (BER), N is the number of bits in the packet. Assuming p_e small, we get

$$p_p \approx p_e N. \qquad (9)$$

To reduce the errors in the transmission of the information packets if they have equal length of the packet N, we need to decrease the value of the bit error p_e as follows from expression (7).

There are known [15] relations for estimating BER when representing the transmission channel by the additive Gaussian white noise model. Therefore, the BER of binary phase-shift keying (BPSK) modulation is

$$p_e = 0.5 erfc\left(\sqrt{\frac{E_b}{N_0}}\right). \qquad (20)$$

In the chirplet expansion, the signal parameters $g(\gamma)= g\ [(l,\ \alpha,\ t,\ \omega)]$ are determined. The received signal is digitized, resulting in a set of $g(\gamma_n)= g[(l_n,\ \alpha_n,\ t_n,\ \omega_n)]$, $n \le N$. And γ_n is the set of possible sampled parameter values form a dictionary D, i.e. $\gamma_n \in D$. Any function $s(t)$ can be represented by a set of atoms $g(\gamma)$. An algorithm that allows searching a suitable combination of data from a dictionary should provide a maximum of the search function

$$s(t) = \sum_{n=1}^{N} s(\gamma_n) g_{c_n} \qquad (31)$$

The parameters $s(t)$ specifying a maximum (10), will determine the maximum approximation to the original signal. In this case, the parameter l_n determines the time domain dilatation, and its reciprocal value is the signal compression in the frequency domain, the ellipse rotation angle α_n corresponds to the linear modulation of the signal center frequency, and the variables t_n, ω_n are the time and frequency of the central part of the signal. It becomes necessary to develop an algorithm for searching $s(\gamma_n)$.

The peculiarity of the algorithm is that a set of parameters is selected from the parameter block dictionary. The algorithm is iterative to provide the best internal signal structure by computing the scalar product of the functions $s(\gamma_n), g_{cn}$. The received value must satisfy the condition

$$|sg_0| = \sup_{\bar{\gamma}} |sg_\gamma| \qquad (42)$$

Further, we compute the remainder term, which at the beginning of the algorithm is equal to the signal itself

$$R_0 = s(t) \qquad (53)$$

and carry out the next steps

$$R_1 = R_0 - |sg_0| g_0 \qquad (64)$$

$$R_2 = R_1 - |sg_1| g_1 = R_0 - |sg_0| g_0 - |sg_1| g_1 \qquad (75)$$

$$R_i = R_{i-1} - |sg_{i-1}| g_{i-1} = R_0 - \sum_{k=1}^{i-1} |sg_k| g_k \qquad (16)$$

...

$$R_i = R_{i-1} - |sg_{i-1}| g_{i-1} = R_0 - \sum_{k=1}^{i-1} |sg_k| g_k \qquad (87)$$

The stopping criterion is the ratio

$$\rho = \frac{\left\| \sum_{k=1}^{n} |sg_k| g_k \right\|}{\|R_n\|} \qquad (98)$$

The higher this ratio value, the worse the chosen decomposition parameters.

4 Experiments

Consider a typical signal LoRa system that is a linear frequency modulated pulse signal. This signal can be obtained, for example, with a voltage-controlled oscillator in the form (1). A useful signal is a packet of pulses with binary modulation. In this case, the logical signal “1”

corresponds to the condition $c > 0$ and the opposite signal, a logical “0”, is obtained if $c < 0$.

In the experiment, a LoRa system transceiver is based on the SX1276 chip to transmit a short message at a frequency of 868 MHz. The receiver is installed on the top floor of the building. The transmitter gradually moves through the building from the top floor to the basement room, which creates interference. In addition, interference is created by mobile communication transmitters, whose antennas are located on the roof of the building. In addition to the signal-to-noise ratio, the reception quality was also determined by determining the number of erroneous bits in a message and measuring the bit error rate (BER). The measurement scheme is shown in Fig. 5.

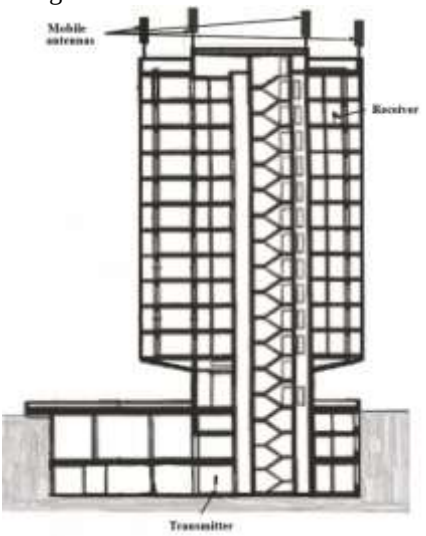


Fig. 5. The scheme of measurement.

The transmitting antennas of mobile operators and Wi-Fi routers that located near the building create jamming with reception. The signals of these devices create an interfering background, which is taken for "white" noise $n(t)$. Crosstalk $w(t)$ is created by multiple re-reflection raying from the interior of the building from the reinforced concrete structures. On each floor of the building, the level of signal and noise is fixed and the quality of the message is controlled. The panoramic receiver selected as the benchmark additionally documents the measurement results. A preliminary analysis of the interference situation presented in Fig. 6.

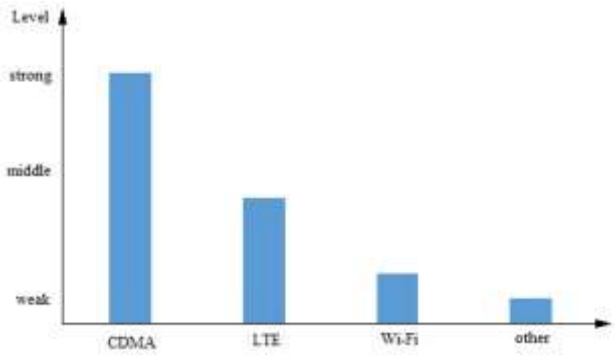


Fig. 6. The radiation intensity.

The results of the measured BER and signal error ratio are presented in Fig. 7.

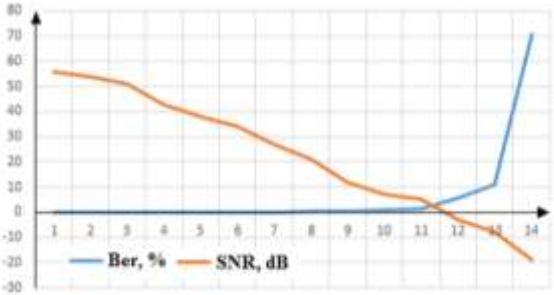


Fig. 7. The measurement of BER and SNR.

The obtained results allow us to represent the vulnerability of the host numerical scale, when the high level of vulnerability corresponds to BER = 10%, SNR = -10 dB, the average level of vulnerability BER = 5%, SNR = -5 dB, weak level BER = 2%, SNR = 2 дБ.

10 Conclusions

Although LoRa technology is relatively new for telecommunication systems, linear frequency modulation signals are used for data transmission, which are considered standard in wireless communication systems over short distances (IEEE 802.15.4a). The complexity of

processing this type of signals is associated with the need for simultaneous time-frequency analysis of the received signal.

Moreover, an improvement in the accuracy of time measurement leads to deterioration in the accuracy of frequency measurement and vice versa, which is explained by the time-frequency uncertainty principle known in radar. The output in this situation is the time-frequency decomposition of the received signal using the matching pursuit algorithm to the message dictionary. The peculiarity of the study is the most suitable use environment-a densely populated district in the city. The result of the study shows that the placement of devices on the surface gives quite good results. Acceptable results are achieved on the 1st and 2nd floor, where BER is about 1% and the signal-to-noise ratio is not worse than 1-2 dB. Future research is planned to focus on the creation of parallel methods for calculation of the set of proposed indicators, the improvement of software, as well as an experimental study of proposed indicators in real conditions.

References

1. Kucherov, D.P.: Control of Computer Network Overload. In: Information Technologies and Security (ITS 2017), pp. 69-75, Kiev, Ukraine, http://ceur-ws.org/Vol-2067/, last accessed 2018/11/21
2. Adelantado, F., Vilajosana, X., Tuset-Peiro, P., Martinez, B., Melià-Seguí, J., Watteyne, T.: Understanding the limits of LoRaWAN. IEEE Communications Magazine, 55 (9), 1 – 7 (2017).
3. Bankov, D., Khorov, E., Lyakhov, A.: Mathematical model of LoRaWAN channel access with capture effect. In: IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), pp. 1 – 5, IEEE, Montreal, QC, Canada (2017).
4. Catherwood, P.A., Little, M., McLaughlin, J.A.D.: Channel characterisation for wearable LoRaWAN monitors. Loughborough Antennas & Propagation Conference (LAPC 2017), pp. 1 – 4, IEEE, Loughborough, UK, (2017).
5. Deng, T., Zhu, J., Nie, Z.: An improved LoRaWAN protocol based on adaptive duty cycle. IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC), pp. 1122 – 1125, IEEE, Chongqing, China (2017)
6. Kucherov, D., Berezkin, A.: Identification approach to determining of radio signal frequency. International Conference on Antenna Theory and Techniques (ICATT), pp. 1 – 4, IEEE, Kyiv, Ukraine (2017).
7. AN1200.22. LoRa™ Modulation Basics. Revision 2, May 2015. 2015 Semtech Corporation, Wireless Sensing and Timing Products Division,

https://www.semtech.com/uploads/documents/ an1200.22.pdf, last accessed 2018/11/21.
8. Mann, S., Haykin, S.: The chirplet transform: physical consideration. IEEE Trans. on Signal Processing, 43(11), 2745 – 2761 (1995).
9. Ashino, R., Nagasw, M., Vaillancourt, R.: Gabor, wavelet and chirplet transforms in the study of pseudodifferential operators. Surikaisekikenkyusho Kokyuroku, 1036 (10098), pp. 23–45, (1997), https://www.osaka-kyoiku.ac.jp/~ashino/pdf/rimsr.pdf, last accessed 2018/11/21.
10. Bultan, A.: A four-parameter atomic decomposition of chirplets. IEEE Trans. on Signal Processing, 47 (3), 731–745 (1999).
11. Yin, Q., Qian, S., Feng, A.: A fast refinement for adaptive Gaussian chirplet decomposition. IEEE Trans. on Signal Processing, 50 (6), 1298 – 1306 (2002).
12. Cui, J., Wong, W., Mann, S.: Time-frequency analysis of visual evoked potentials using chirplet transform. Electronics Letters, 41 (4), 217 – 218 (2005).
13. Wang, X., Fei, M., Li, X.: Performance of chirp spread spectrum in wireless communication systems. In: 11th IEEE Singapore International Conference on Communication Systems (SICCS), pp. 466–469, IEEE, Guangzhou, China (2008).
14. Bultan, A., Akansu, A.N.: A novel time-frequency exciser in spread spectrum communications for chirp-like interference. In: IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP '98), pp. 3265 – 3268, IEEE, Seattle, WA, USA (1998).
15. Shin, Y.S., Jeon, J-J. Pseudo Wigner-Ville time-frequency distribution and its application to machinery condition monito

ПЕРЕДОБРОБКА І АНАЛІЗ НАБОРІВ ЗОВНІШНІХ ДАНИХ В ЕЛЕКТРОННОМУ ДОКУМЕНТООБІГУ КРИТИЧНОЇ ІНФРАСТРУКТУРИ

Кузьмичов А.І.
Інститут проблем реєстрації інформації НАН України
Вул. М. Шпака, 2, 03113 Київ, Україна

Електронний документообіг – реально діюча інформаційно-аналітична складова критичної інфраструктури, комплексу із розвиненого програмно-технічного інструментарію, організаційних систем характерних конфігурацій та джерел зовнішніх даних різноманітних форматів, що забезпечує організацію взаємодії інформаційних потоків (документів), функціонування та розвиток засобів інформаційної взаємодії.

***Ключові слова**: електронний документообіг, критична інфраструктура, передобробка і аналіз, набори зовнішніх даних, dataset, data engineering, data preprocessing and analysis.*

Вступ

Вміст критичної інфраструктури [1] змушує застосовувати до її поточного електронного документообігу специфічні вимоги, задоволення яких змушує представляти процес електронного документообігу критичної інфраструктури (ЕДКІ) як процес поглибленого аналізу даних (дейтамайнінгу, майнінгу наборів даних, [2]), де послідовно реалізуються автоматизовані процедури перетворень: набори даних великих розмірів → інформація → знання.

Цей процес – предмет міждисциплінарної області технологій, навичок і знань – комп'ютерних наук, математики і статистики, штучного інтелекту і машинного навчання, що має пряме відношення до проблем інформаційної безпеки [3].

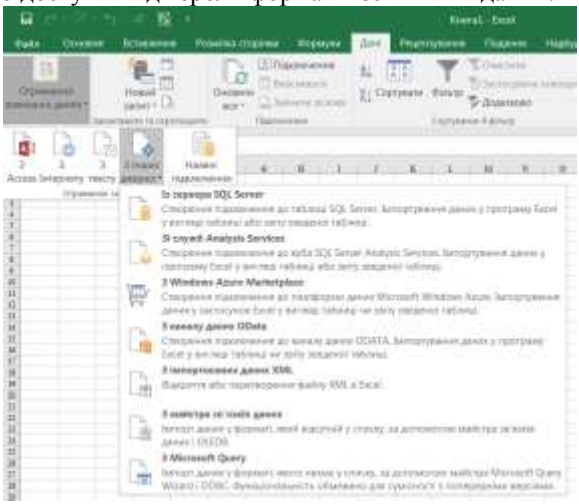
Але, як будь-який потужний інструмент чи засіб, їх корисність залежить від того, у чиїх руках він знаходиться й з якою метою використовується, бо за його допомогою можна виявити шукані витоки негативу чи зробити неочікувану серйозну шкоду. На цінний результат можна сподіватися за умови якісних початкових даних, які отримують ззовні і попередньо обробляють розвиненими інструментальними засобами на потужній операційній платформі.

Майнінг даних в Excel

ЕДКІ – поширена область інформаційно-аналітичної діяльності у розмаїтті критичних інфраструктур: це увесь бізнес, науково-дослідна і проектно-конструкторська робота, спеціальні служби, ситуаційні центри, установи державного управління тощо. Тут діють наукоємні (інтелектуальні, high-tech) інформаційні системи (типу розгалужених банківських чи пошукових систем), які, потенційно, піддаються атакам (вторгненням) ззовні та/чи з середини, бо містять зосереджені джерела даних чи інформацію про них, корпоративні «знання» (ноу-хау) й усю необхідну інформацію (у т. ч., відкриті дані) для ефективної діяльності. Зокрема, різновидами вторгнення з середини інфраструктури - витік інформації, де джерелом даних є персонал, власний чи організацій-партнерів, або утворення аномалій певного класу.

Ця поширеність визначає популярність доступних офісних програмних продуктів новітніх версій, де, відповідаючи на запит практики, розробники завчасно передбачили отримування даних із різноманітних зовнішніх джерел для подальшої обробки і аналізу у цьому ж операційному середовищі.

У першу чергу, це роблять власники офісних пакетів, та ж компанія-власник Microsoft Corp., створюючи окремий продукт типу MS SQL Server Data Mining чи модифікуючи існуючі програмні продукти масового користування. Ось, скажімо, пропозиції Excel 2016 щодо доступних джерел і форматів зовнішніх даних:



Компанії-партнери для популярного табличного процесора Excel розробляють зручні програми-надбудови (*add-ins*) із наборами досить потужних аналітичних інструментів.

Ось одна з них від компанії Frontline Inc. (добре відомої аналітикам надбудовою Excel Solver, Поиск решения) – це надбудова XLMiner [4, 5], інструменти якої застосовують апарат імітаційного моделювання та математичного програмування:

.

Тут, отримавши зовнішні дані (Get Data), можна їх дослідити, почистити і побудувати з них вибірки (Explore, Transform), виконати кластеризацію (Cluster) та класифікацію (Classify) для виявлення аномалій, провести якісний аналіз часових рядів, популярний в поточній бізнес-практиці (Time Series), зробити прогноз (Predict) тощо, і цих інструментів, як свідчить світова практика, буває достатньо для попереднього аналізу отриманих зовнішніх даних і, зрозуміло, для якісної навчальної роботи й підвищення кваліфікації персоналу.

Але цим основним діям передує важливий етап передобробки (preprocessing) отриманих «сирих» даних, де, зокрема, застосовуються модифіковані стандартні інструменти Excel:

Тут для роботи з даними великих розмірів корисними є інструменти: *Текст за стовпцями* (перетворення з форматів .txt, .csv в формати .xls, .xlsx), *Видалення повторів*, *Перевірка даних*, виклик *Power Pivot* при роботі з великими даними тощо.

Із назви і вмісту процесу і техніки дейтамайнінгу (або коротко, майнінгу, видобування знань) видно, що мова йде про новітню версію звичайної практики усіх часів: дослідник, маючи отримані дані, буває, непростими і недешевими способами (як от, у наші дні, ґрунт Місяця), наступним їх поглибленим вивченням за

можливостями існуючих на той час інструментальних засобів має знайти щось характерне і невидиме, insight, «знання», аби зрозуміти природу об'єкту, представленому наявними даними.

Так було завжди, в наші часи є «великі» і «сирі» дані, що автоматично генеруються і зберігаються у вигляді наборів даних неозорого розміру, що вимагає для роботи з ними серйозних комп'ютерних засобів – процесорів, пам’яті і наукоємних спеціальних програм (майнерів).

Рядовий користувач знає, що сторінка Excel має десь 1 млн. рядків і 16 тисяч стовпців, чув, що якщо треба більше, застосовують програми-надбудови для масивів у кілька мільйонів записів як от MS Power Pivot:

Але, розпочавши роботу із зовнішніми чи відкритими наборами даних, імпортувавши файл табличних даних із десятків чи сотень тисяч записів, можна неочікувано здивуватися: мій ПК, який, мені здається, має фантастичні параметри, навіть виконуючи нескладну роботу типу фільтрації змушує якийсь час чекати її завершення, і це означає, що проблематика майнінгу великих даних під силу лише потужним апаратним засобам й, до того ж, недешевому софтверу.

Проблематика аналітики даних визначила специфіку й кадрового забезпечення цієї діяльності, як з’ясувалося із досвіду, досить тривалою і нестандартною, навіть критичною (наприклад, в соціології, епідеміології чи в аналізі медичних текстів) є передобробка і аналіз сирих наборів даних, для цього напряму навіть виокремлено вкрай дефіцитну спеціальність широкого профілю Data Engineering. І це зрозуміло, адже за виразом «Що посієш, те й пожнеш», майбутній результат (вихід) прямо залежить від входу, від якості попередньої обробки, перевірки і зрозумілості початкових даних, зауважимо, неохопленого очима масиву (і, буває, нечитабельного чи неохайно оформленого), як ось такий (про аварії в авіації):

	AIRPORT_CITY	AIRPORT_STATE	AIRPORT_COUNTRY	ICAO	LCHG_DATE	RECORD_STATUS	Record_End
28923	SHENYANG	CH	Y	04/17/2003	I		
28924	MUNI	ONT,AIRPORT	BINACE	MUNI	ONT	CA	12.03.1990
28925	PARRY	NWT,AIRPORT	CAPE	PARRY	NWT	CA	12.03.1990
28926	FALLS	NFLD,AIRPORT	CHURCHILL	FALLS	NFLD	CA	01/31/1990
28927	ROCK	ZUNI	PUEBLO	NM	07/25/2018	I	
28928	SHUANGLIU	INTERNATIONAL	AIRPORT	CHENGDU	CN	Y	06/14/2012
28929	ALTA,AIRPORT	EDMONTON/VILLENEU	ALTA	CA	01/31/1990	A	
28930	BROCHET	MAN,AIRPORT	LAC	BROCHET	MAN	CA	01/31/1990
28931	LAKE	SASK,AIRPORT	WOLLASTON	LAKE	SASK	CA	01/31/1990
28932	31	WEST	CASIP	WAINWRIGHT	ALTA	CA	06.02.1989
28933	DWOPU	INTERNATIONAL	AIRPORT	URUMQI	CN	Y	06.08.2012
28934	DONGTAO	AIRPORT	JIANUSI	CH	Y	09/26/2016	A
28935	SHENYANG	CH	Y	04/17/2003	A		
28936	MUNI	ZANESVILLE	OH	08.02.1989	A		

Вже далі «чисті» дані у форматі, скажімо, табличної бази даних передаються для кінцевої і основної роботи «шахтарю» – добре освіченому спеціалісту вузького профілю (Data Scientist) саме для виявлення і видобування «знань», де надто важлива співпраця, скажімо, IT-спеціаліста із фахівцями певної області досліджень, аби коректно за наявними даними визначити відповідну інформацію. Тому й кажуть, що аналіз даних це технологія не про числа чи тексти, а про запитання і відповіді, аби сформувати необхідне розуміння ситуації для наступної перевірки припущень щодо шуканих «знань».

Тож маємо усвідомлювати, що в аналізі «брудних» даних дата-аналітика може цікавити саме «бруд», наприклад, пропуски чи хибні значення певних показників, зроблені із якихось причин. Саме тому, автоматизована передобробка сирих даних великого розміру – предмет серйозних наукових та прикладних досліджень [6], де визначають типи хибних даних (недійсні, неоднозначні, викиди, пропуски, помилки), аби їх вчасно виявити і врахувати для налаштування програм-майнерів, що мають для цього засоби статистичного моделювання, вибіркового дослідження та візуалізації.

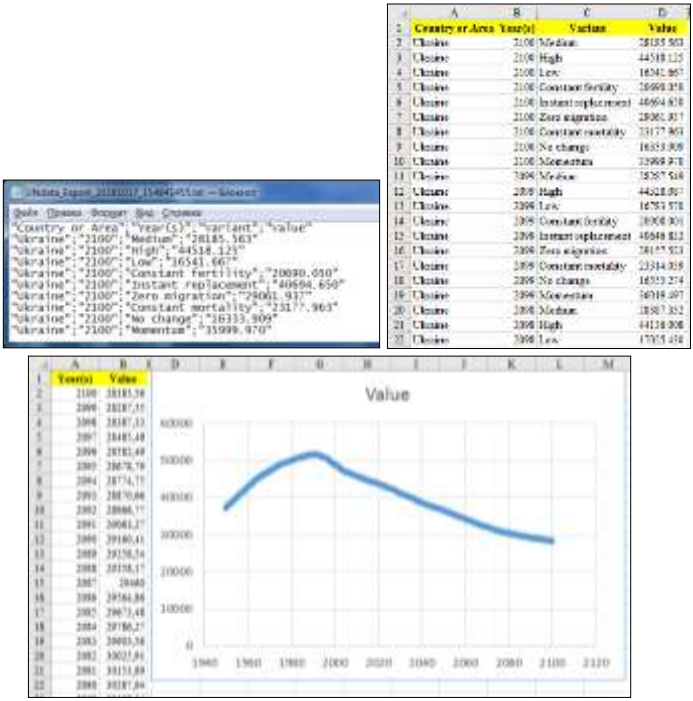
Набори зовнішніх і відкритих даних

Увесь процес майнінгу даних базується на наборах зовнішніх даних, тож надзвичайно важлива культура формування цих наборів, що визначає їх якість й успішність усього процесу.

«Ідеальними» наборами зовнішніх даних слід вважати такі, які зразу ж після отримання (імпорту) можна обробляти наявними інструментами, наприклад, відшукувати прості аномалії: повтори (дублікати) записів чи пропуски в записах.

Приклад 1. Набір із 1359 записів про населення України на період 1950-2100 рр. у текстовому форматі представлено на сайті ООН data.un.org:

Після імпорту перетворенням txt → xlsx можна зразу виконати сортування чи фільтрацію таблиці, а після заміни десяткових точок в числах комами, побудували точкову діаграму (вибравши опцію Medium):



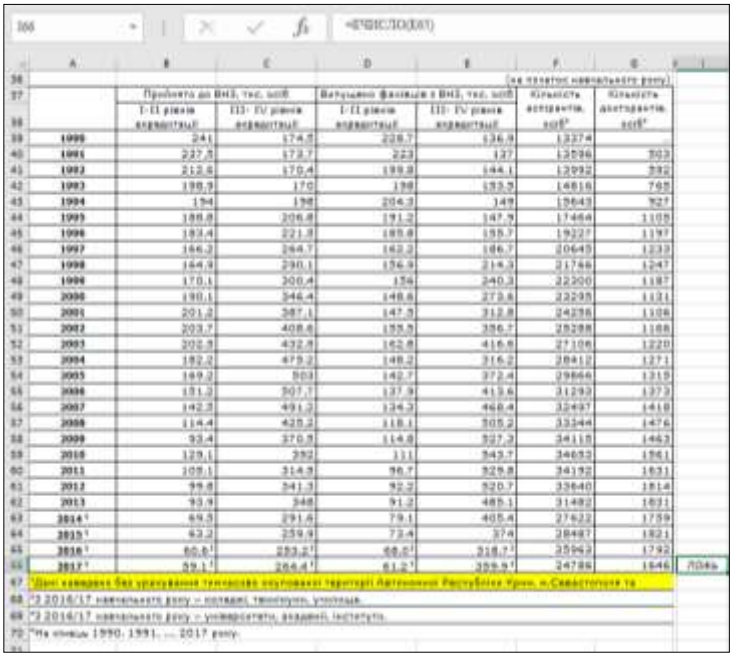
Приклад 2. Дані із посиланнями. Маємо для порівняння два набори даних із посиланнями, де

файл (сайт ООН) містить посилання в окремому стовпці (Value Footnotes), у кожному полі дані одного типу, таблична база даних

ізольована (рядком 401) і документ можна обробляти зразу після імпорту:

.

Файл (сайт Укрстату, ukrstat.gov.ua) розроблений, скоріше за все, лише для друку і для обробки не готовий: містить посилання безпосередньо в значеннях даних (клітинки D65, D66 та ін.), що робить їх хибними (не числами), таблиця не ізольована (рядки 67 ÷ 70 з посиланнями), тож передобробка необхідна:



Отже, досвідчені практики рекомендують скептично відноситися до сирих даних, бо то ще «чорна скриня» і це зрозуміло, тому перший крок – просто зрозуміти те, що ми отримали: як визначені поля, дані в них якого типу, у якому вимірі й форматі представлені, адже в Excel це лише: формули, числа (включаючи дати) й текст.

Приклад 3. Пропуски значень, невідповідність значень типу змінної (полю).

Списки депутатів Верховної ради України (сайт ВРУ) мають статус відкритих даних, їх формують у спеціальному департаменті в структурі ВРУ і, очікується, що це зразки оформлення наборів зовнішніх даних. Але ось один з них (усього до 500 записів) – список депутатів діючого 8-го скликання містить багато пропусків значень полів, «комбінованих» значень теж багато як от: для депутата Котвіцького (рядок 209) у полі опису партії (party_text) чомусь вказано ще про його участь у ПАТ «Інвестор»:

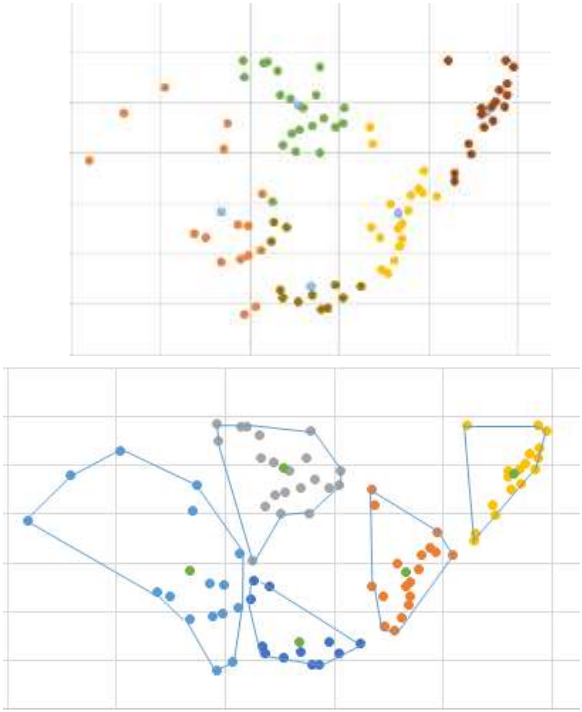
або: у полі district_text (довжиною у 123 символи) для опису виборчого округу вказано два різних тексти (округ, партія) і число (11), хоча для них є спеціальні поля: district_name, party_name та num_in_party:

Напрошується висновок, що розробники закордонних джерел зовнішніх і відкритих даних досить відповідально відносяться до формування наборів даних, публікуючи їх не лише в текстовому форматі для імпорту у різні середовища, а й у вигляді табличних баз даних, готових для проведення попереднього аналізу в Excel.

Кластеризація

Задача [7-9]. Для обслуговування *n* об'єктів (клієнтів) треба знайти місця (координати) оптимального розміщення *K* серверів (значення *K* задано чи відшукується) за критерієм мінсум (мінімізація сум відстаней між клієнтами і серверами), група клієнтів, що обслуговується певним сервером, утворює кластер, найкраще це видно на точковій діаграмі, якщо клієнтів задати

зваженими координатами (x, y, m), K = 5, n = 84 (населені пункти Криму):



Це досить складна й важка задача нелінійного програмування з невідомими (0,1)-типу, існуючі програмні засоби дозволяють знайти локальний оптимум з-за нелінійної функції обчислення відстані між двома об'єктами, для двох точок a(x_a, y_a) та b(x_b, y_b) це:

$$d(a,b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2} \ .$$

Ці засоби реалізують два характерних типи обчислювальних алгоритмів: точні, але складні градієнтного типу, й алгоритми еволюційного програмування, які дозволяють отримати наближений оптимум шляхом досить тривалого направленого перебору варіантів. Складнощі додаються умовою, щоби шукані невідомі приймали одне з двох значень, 0 чи 1, бо клієнт має обслуговуватися одним сервером, існують серйозні вимоги щодо розмірів моделі (числа шуканих невідомих та обмежень на їх значення).

Засоби кластерного аналізу у складі майнерів даних застосовують для виявлення аномалій (помилок, відхилень, виключень, зломів) в наборах даних великих розмірів, використовуються засоби імітаційного моделювання (генератори випадкових чисел) та алгоритми еволюційного програмування: випадковим чином визначаються стартові позиції кластерів, які надалі в ітераційному процесі уточнюються розрахунками відповідних відстаней.

Результати (K = 5, n = 84):

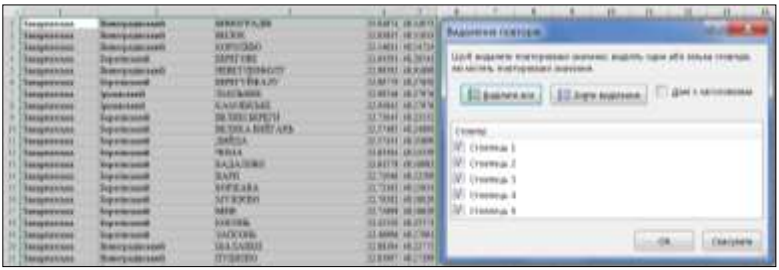
Cluster	x	y
Cluster 1	33,735645	44,49144389
Cluster 2	34,13173611	44,49066889
Cluster 3	33,905773	44,5970325
Cluster 4	34,32841706	44,59131294
Cluster 5	33,93541	44,41894091

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	0	0,396091869	0,200231094	0,601126061	0,212515264
Cluster 2	0,396091869	0	0,249744961	0,220935784	0,209018767
Cluster 3	0,200231094	0,249744961	0	0,422682758	0,180540761
Cluster 4	0,601126061	0,220935784	0,422682758	0	0,429146439
Cluster 5	0,212515264	0,209018767	0,180540761	0,429146439	0

Cluster	Size	Average Distance
Cluster 1	16	0,105443474
Cluster 2	18	0,050892334
Cluster 3	21	0,06579086
Cluster 4	17	0,047717892
Cluster 5	12	0,064417055
Total	84	0,066297315

Record ID	Cluster	Dist.Cluster-1	Dist.Cluster-2	Dist.Cluster-3	Dist.Cluster-4	Dist.Cluster-5
Record 18	1	0,044714268	0,433751195	0,244913534	0,642035508	0,241430551
Record 24	1	0,178332332	0,538995547	0,298982197	0,721466258	0,382340403
Record 49	1	0,061672608	0,336145688	0,16205234	0,544512012	0,1515526
Record 53	1	0,062492372	0,357395322	0,199166751	0,570256458	0,160228635
Record 55	1	0,071989908	0,340528838	0,185653376	0,553271173	0,144788245

Точкові аномалії – записи бази даних типу дублікатів порівняно легко виявляються не лише існуючими майнерами, а й інструментом *Видалення повторів* Excel:



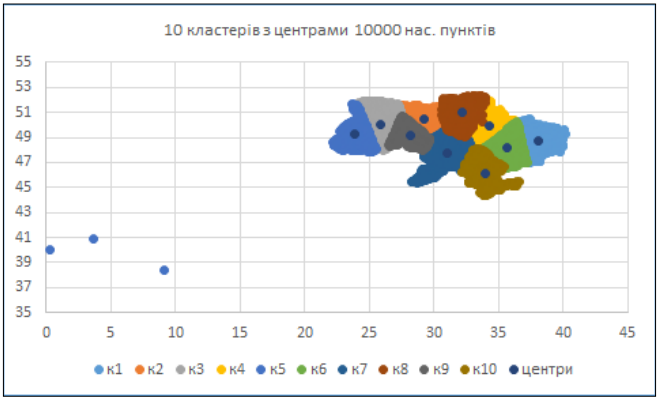
Умовні (контекстуальні) аномалії визначаються значеннями змінних (атрибутів) записів, скажімо, координати об'єкту (широта, довгота) цілком «нормальні» щодо типу і формату, зате аномальні за умови контексту – області досліджуваного простору чи території.

Колективні аномалії утворює група умовних аномалій однакових властивостей.

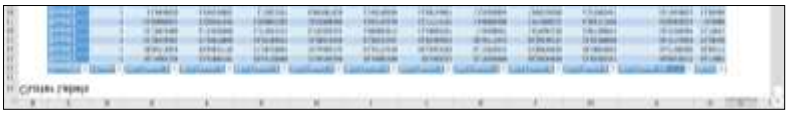
Визначені аномалії можуть утворити окремий кластер чи бути часткою «нормального» кластеру, виникає задача виявити їх певним чином. Для об'єктів із заданими координатами можна візуалізувати кластери, їх центри (центроїди) та аномалії.

Приклад 4. Для розглянутого вище набору даних про майже 30 тисяч записів про населені пункти України в XLMiner сформована сертифікована без повторень вибірка розміром 10 тис. записів (згідно обмежень пробної версії), для якої інструментом *Data Mining → Cluster → K-Means Clustering*, $K = 10$ (максимум), визначено кластери без та з аномаліями із відповідними показниками: координати центрів 10 кластерів, матриця відстаней між цими центрами та відстаней кожного запису до кожного центру, належність кожного запису відповідному кластеру.

В Excel за координатами 10 тис. населених пунктів України та трьох умовних аномальних (за їх координатами) населених пунктів (Мадрид, Валенсія, Лісабон), центрів 10 кластерів побудовано точкові діаграми, де явно видно три умовних аномальних пункти у складі кластеру 5 (населені пункти Західної України):



Найчастіше побудувати точкову діаграму неможливо, тоді аналізом таблиці відстаней кожного об'єкту (запису) від центрів кластерів визначають умовні аномалії як викиди (у прикладі записи 1 ÷ 3 для трьох неукраїнських міст найвіддаленіші від усіх центрів), які мають привернути увагу аналітика даних:



Висновок

За наведеними ілюстративними прикладами показано, що інструментар майнінгу даних є потужним і ефективним засобом для убезпеченого електронного документообігу критичної інфраструктури. Для якісного здійснення цього виду інформаційно-

аналітичної діяльності необхідно звернути увагу на важливість етапу передобробки сирих наборів зовнішніх даних, від чого залежить якість виконання основної роботи із виявлення несанкціонованих втручань в інформаційне середовище. Цей етап можна здійснити лише маючи розвинений апаратно-програмний комплекс підтримки вказаного виду діяльності. Також показана необхідність високої культури на етапі формування зовнішніх даних, які мають бути корисною базою даних функціонування критичної інформаційної інфраструктури.

Література

1. Про критичну інфраструктуру та її захист. Проект закону .України (2018)/
2. Aggarval C. Data Mining: The Textbook. – Springer. – 2015. – 746 p.
3. Niranjan A. Security in Data Mining – A Comprehensive Survey, in Global. Jorn. of CS & Tech. – 2016, v. 16. – 50-72 p.
4. Analytic Solver Data Mining. User Guide. V. 2018. www.solver.com
5. Shmueli G. Data Mining for Business Analytics. Concepts, Techniques, and Applications with XLMiner/ G. Shmueli, P. Bruce, N. Patel. – Wiley, 2016. – 549 p.
6. Cuesta H. Practical Data Analysis. Transform, model, and visualize your data through hands-on projects, developed in open source tools. – Packt Publ., 2013. – 360 p.
7. Кузьмичов А. І. Оптимізаційні методи і моделі. Моделювання засобами MS Excel. – ІПРІ НАНУ, 2017. – 428 с. (електронний ресурс)
8. Daskin M. Network and Discrete Location. Models, Algorithms, and Applications, 2-ed. – Wiley, 2013. – 535 p.
9. Дюран Б. Кластерный анализ/ Б. Дюран, П. Оделл. Пер. с англ . М.: «Статистика», 1977. – 128 с.

INTEGRATED KNOWLEDGE DESCRIPTION MODEL FOR ANALYTIC ACTIVITIES

Senchenko V.R.
Institute for Information Recording of National Academy of Sciences of Ukraine, Kyiv, Ukraine

For any intellectual system based on knowledge, the very construction of its model of knowledge base is the most complicated and responsible process. Analysis of trends in the development of analytical systems shows that modern systems need to combine different methods of describing knowledge, since no single method can fully provide a description of the model of knowledge of the real analytical system. Knowledge models has indirectly related to the type of input date (structured, unstructured, fictitious, text, hypertext and other types of data) that form the context for tasks solved in the analytic system. Has proposed an approach for describing the integrated knowledge representation model. This approach based on the Semantic Web paradigm, which provides mechanisms for integrating various knowledge description models through used metadata describing different data models and their semantic connectivity based on graph description of complex dependencies. The proposed integrated model of description of knowledge for carrying out analytical activity combines object-oriented approaches with semantic representations on knowledge of the subject domain. This allows using different models of knowledge in one analytical system (ontological, network, relational, production) and data that can be located in a distributed environment, including WWW space.

Key words: *Knowledge base, knowledge base model, output logic, analytic activity, ontology, scenario*

Introduction

For any intellectual system based on knowledge the very construction of its knowledge base model is the most complicated and responsible process. First, it is necessary clearly distinguish the nature of knowledge about the subject domain (especially if it relates to multi-factor systems), that is, which model best describes knowledge of the subject domain

(SD): relational, framing, object-oriented, semantic-network, hypertext, ontological or other models.

Thus, semantic networks [1, 2] represent an informational model of the domain, which has the form of a directed graph whose vertices correspond to the objects of the domain, and the edges define relations between them. Objects can be concepts, events, properties, processes used to store information in terms of objects and relations between them. Knowledge bases of the analytical system should also contain specific knowledge, which provided in the form of instances objects (classes) and the links between them or the restrictions given on the values of attributes of instances of concepts.

Production systems [3] are considered as a means of representing experts' knowledge in the form of rules of the form "IF – THEN" and performing logical inference. The process of extraction of new knowledge is a search by rules based on samples - existing facts, events that determine the current state of PR and are contained in the knowledge base.

Frame models [4] represent an object-oriented approach and serve both for raising the level of knowledge representation, and for ensuring the possibility of sharing the use of declarative and procedural knowledge. The first type includes knowledge describing the system of problems of the SD, including the division of tasks into subtasks and the relationship of subtasks with the methods of their solution, the second – knowledge, representing both methods of problem solving, and specific output algorithms.

For the description of the object-oriented model of knowledge the basic essence of the SD (concepts and objects) must be determined, as well as knowledge of how these entities are interconnected [5, 6]. That is, this model also defines the knowledge of relationships that directly link the concept of various types of relationships between the notions of SD, which are described both on the logical and functional levels.

Should be reminded that for many SDs there is often a need to use knowledge that has inaccurate values of attributes - "fuzzy logic" [7]. Consequently, for such knowledge, it is necessary to use methods and procedures for describing the fuzzy values. Therefore, an adequate means of formalizing the ontology can be models based on linguistic variables, fuzzy sets, fuzzy relations, fuzzy graphs and fuzzy trees, fuzzy restrictions. That is, in order to represent fuzzy knowledge in the SDs model, it is necessary to provide mechanisms for installing underexposed values in the attributes, as well as assigning restrictions to the attribute values that bind such attributes

Existing CASE tools like Rational Rose, ERwin, Silverrun, BPwin, S-Designor, Visible Analyst, Visual, CASEAnalyst and others focuses on accelerating the process of designing the structure and code of the software systems. With their help, you can really build business process models in the form of data flow diagrams, perform conceptual modeling of data, and relatively quickly create software applications for data analysis. Some CASE tools (for example, Silverrun) even allow for the creation of simplified knowledge models for NoSQL databases, but, in general, with their help, it is impossible to describe an integrated model of knowledge that most modern analytical systems require.

The research task

The analysis of modern trends shows that when creating applied analytical systems, it often becomes necessary to combine different methods of describing knowledge, since no single method can fully provide a description of the model of knowledge of the real analytical system. This is because knowledge models also indirectly related to the type of input data (structured, unstructured, fuzzy, text, hypertext and other types of data) that form the context for problems solved in the analytic system.

Therefore, it is desirable that the integrated knowledge model should take into account not only the methods of representing knowledge of different models (ontological, network, production, etc.), but also the means of forming logical rules and processes for the emergence of new knowledge.

Thereby, there is an objective need to develop an approach for describing an integrated knowledge model that combines object-oriented methods for describing data models with their semantic representations that are sufficient for the interaction of different knowledge models within the framework of the created analytical system.

The main material

An approach to describing an integrated knowledge model has based on the Semantic Web paradigm [2]. Semantic Web's paradigm involves the dissemination of knowledge (metadata) in languages specially designed to work with different data formats: Framework Description (RDF), Web Ontology Language (OWL) and Extensible Markup Language (XML) [7]. HTML describes documents and links between them. That is, the Semantic Web paradigm provides the necessary

mechanisms for integrating various knowledge description models using metadata describing different data models and their semantic connectivity based on graph description of complex dependencies - RDF graphs and XML.

Consequently, the integrated knowledge model - $InM(KB)$, based on the Semantic Web paradigm has described by a mathematical expression:

$$InM(KB) =< Ont(SD), SNet, FScen, PRul, PCase, PRes >$$

where $Ont(SD)$ – applied ontology, which describes the main entities (concepts and relations) of the subject area in the form of classes of objects, instances of classes, their properties and relations between classes and properties, including a description of the information resources necessary for carrying out analytical activity:

$SNet$ – semantic network that describes the properties of SD objects and information resources, as well as the relationships between components of the integrated model, using network methods for describing components in the form of a graph model;

$FScen$ – model of analytical scenarios that describes the analytical functions and procedures inherent in PR. Based on the components of the functional network, executive scenarios of analytical activity are formed for carrying out operations on instances of the classes of applied ontology. Model scripts analytic activity - takes into account the real limitations on the values of the attributes of semantic network objects - $SNet$;

$PRul$ – production rule model – describes inference rules in terms of classes and relations for performing operations on instances of applied ontology classes $Ont(SD)$. The model of production rules based on axioms that define consistent statements for deduction based on descriptive logic, or a set of precedents for fuzzy output.

$PRes$ – model of the mechanisms for the output of new knowledge (software tools for logical inference) and data processing For modern analytical systems, this model should contain appropriate mechanisms for both clear conclusions (based on rules and axioms) and output mechanisms based on fuzzy rules that manipulate linguistic variables

$PCase$ – executive software model of analytical activities, consists of individual software blocks and components designed to implement scenarios, a variety of relevant analytical functions, data processing procedures and the output of new knowledge.

As a basis for an integrated knowledge model – $InM(KB)$ proposed to use the model of applied ontology - $Ont(SD)$, which is the main tools of describing both the subject area and the software that implements it.

Such an ontology defines the main entities (classes of objects, properties and relations between them), which are then used by other means of the model. Formally, subject domain ontology represented as a mathematical expression:

$$Ont(SD) = < C^{(SD)}, R^{(In)}, Rb^{(C)}, T^{(SD)}, Ac^{(C)}, Cntr, Rul^{(S)} >$$

where $C^{(SD)} = \{C^{(SD1)}, \dots C^{(SDm)}\}$ – a lot of classes, a description of the basic understanding of the subject domain and a software application, which is a real-life analytical system;

$R^{(In)}$ де $R^{(In)} \subseteq C^{(SD)} \times C^{(SD1)}$ – strict partial order on the set of classes $C^{(SD)}$, that sets the success rate;

$Rb^{(C)}$ де $Rb^{(C)} = Rb^{(C1)}, \dots, Rb^{(Cn)}$, in turn $R^{(Ci)} \subseteq C^{(SD)} \times C^{(SD1)}$ – the finite set of binary relations that are established between two elements of the set given on the classes of ontology - $C^{(SD)}$;

$T^{(SD)} = \{ T^{(SD1)}, \dots, T^{(SDt)}\}$ – types of data that characterize data processing features for a SD;

$Ac^{(C)} = \{ a_1^{(Ci)}, \dots , a_r^{(Ci)}\}$ – a set of attributes describing the properties of classes - $C^{(SD)}$ the subject domain;

$Cntr$ – the rrestriction's set on class attributes $Ac^{(C)}$ been written logical expressions of the form:

$Cntr(Cn_{i1}, \dots, Cn_{iw})$ де $Cn_{ik} \in Ac^{(Ci)}$ або $Cn_{ik} \in T^{(SDi)}$, i.e. Cn_{ik} can be either the name of the attribute, or the constant used to determine the deterministic terms of the knowledge output through the corresponding programs.

$Rul^{(S)}$ – subset of rules for logical output for obtaining new knowledge in SD due to logic output machine - $PRes$.

Thus, an applied ontology $Ont(SD)$ defines a comprehensive description of the concepts and relations of both the subject domain and the software that implements it. This description carried out in the form of objects and relations classes that encapsulate semantic properties and limitations on their attributes.

Such a description actually determines the structure of the semantic network, which describes the information sources, the taxonomy of the main concepts of subject domain. This description also relates them to the scenario model and descriptive logic of processing, which also provides knowledge of the subject domain, as well as knowledge of the functions that are appropriate to perform in solving concrete analytical tasks.

Iti also has be emphasized that with the help of an ontological model, the entities of SD defined, in the terms of which the production rules of the logical conclusion of new knowledge has described.

The main advantage an ontological model in front of other models of knowledge is the ability to describe the hierarchy of "general-private" relationships for any class of ontology – $C^{(SD)}$. This advantage greatly simplifies the process of describing the PRO by using the inheritance mechanisms of properties (including attributes, relationships and constraints) of the higher classes of lower ones.

Thanks to the inclusion to the integrated model of the apparatus of the linguistic types of data, all the simple types of data included in the set of data types - $T^{(SD)}$, have linguistic extensions. This allows you to define the linguistic values of data in the form of an interval of allowable values of $Cntr$ and then to operate with them. In particular, set attributes of objects on tological model certain values that contribute to more accurate logical output.

In this sense, the functional dependences $a_j \in Ac^{(C)}$ between the attributes of entities can be given in the descriptions of classes of objects and relations in the form of restrictions $Cntr_i \in Cntr$, which bind the values of the attributes of entities (that is, in the descriptions of binary relations, restrictions are set on the attributes of their arguments.) Such constraints are the finite set of logical expressions that connect the logical operations of the value of the attributes of object objects.

To describe the rules for logical derivation, $PRul$ is used as a model of statements of facts containing the ontology of the subject area, presented in the form of an axiom, as a model of statements in the form of "IF – THEN". In both cases, a semantic network of concepts is used which are defined in terms of the subject area.

Functional or scenario model - $FScen$ occupies a special place in the integrated model. It consists of typical data and information processing scenarios presented by different types (structured, unstructured, poorly structured, textual, Web) for obtaining new knowledge. The scenario model links the algorithms of the software implementation of the analytic system – $PCase$ with the model of production rules. If the implementation of the proposed means is complicated by significant costs or ineffective, it expected to supplement the model with more effective data processing methods.

The interconnections of the integrated model of knowledge for carrying out analytical activity are shown in Figure 1.

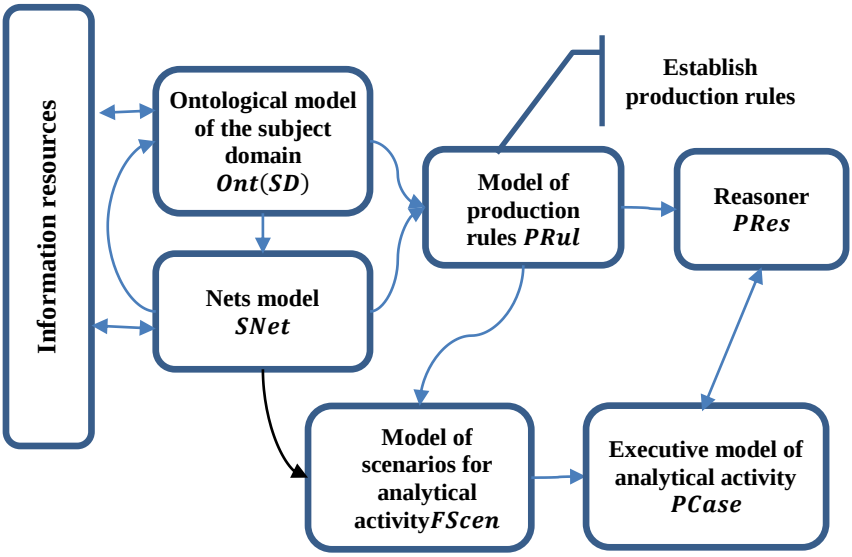


Fig. 1. Interconnections in an integrated model of knowledge for carrying out analytical activity

The integrated model allows to represent the task of the analytical system using an ontology model – $Ont(SD)$. Model describes the types of tasks and methods for their solution in the form of typical scenarios of analytical activity – $FScen$. Besides model using the semantic network – $SNet$, that divided of tasks into subtasks and the relationship of tasks with the methods and information resources needed to solve them is given.

At the same time, the solution methods themselves can be represented by production rules combined in the script – $FScen$. To do this, the model includes the means of the task of production rules «Establish production rules» and their management – $PRul$, as well as activation operators of rules that can be combined with executive controls and executed using logic output machines – $PRes$.

To implement the proposed approach, the most effective tool for developing an ontological model is the Ontology Editor Protégé 5 developed by Stanford University [9]. The editor Protégé 5 relates to visual editors - Visual methods for designing ontologies help to quickly and fully understand the domain knowledge structure, which is especially valuable for researchers working in a new subject area.

Protégé 5 allows you to support all phases of the ontology life cycle in accordance with the requirements of ISO / IEC 15288: 2002 [10] - from the development of a semantic network and the creation of a knowledge base on its basis, to the formation of user requests to these databases for the purpose of obtaining knowledge.

The proposed integrated model of description of knowledge for carrying out analytical activity combines object-oriented approaches with semantic representations on knowledge of the subject domain. This allows using different models of knowledge in one analytical system (ontological, network, relational, production) and distributed data, including WWW.

In this case, to universalize the description of knowledge been used:

- Concepts of SD and software has represented by the classes of ontology objects that encapsulate semantic properties and limitations on attributes;
- objects in the semantic network determine the information resources needed to derive new knowledge based on interaction with the scenario model and rules of logical output;
- all the necessary processes for the output and processing of information are implemented by the system of production rules, which is connected with the semantic network and scenarios for the implementation of analytical functions and procedures that carry out the tasks of the analytical system;
- The implementation of scripts for the execution of functions and carried out procedures through the programming tools and the mechanism for the output / processing of knowledge and data and specialized programs (logic machines).

References

1. Martin D., Paolucci M., McIlraith S., Burstein M., McDermott, D., McGuinness D., .& Srinivasan, N. (2004). Bringing semantics to web services: The OWL-S approach. In Semantic Web Services and Web Process Composition (pp. 26-42). Springer Berlin Heidelberg.
2. What is Semantic Technology?: URL: https://ontotext.com/knowledgehub/fundamentals/semantic-web-technology/
3. Klahr, D., Langley, P. and Neches, R. (1987). Production System Models of Learning and Development. Cambridge, Mass.: The MIT Press.
4. Frame language https://en.wikipedia.org/wiki/Frame_language
5. Object Management Group / Modeling Specifications: http://www.omg.org/spec/#M&M

6. Object-data-model. URL: http://www.gartner.com/it-glossary/object-data-model/
7. Zade, L. The concept of a linguistic variable and its application to making approximate decisions [Text] / L. Zade. –M .: Mir, 1976. - 166 p.
8. OWL 2 Web Ontology Language Document Overview (Second Edition) W3C Recommendation 11 December 2012: URL: http://www.w3.org/TR/owl2-overview/
9. Protégé - Free, open-source ontology editor and framework for building intelligent systems: URL: http://protege.stanford.edu
10. ISO/IEC 15288:2002 «System engineering — System life cycle processes»

ANALYTICAL TECHNOLOGIES FOR CLIENTS' PREFERENCES ANALYZING WITH INCOMPLETE DATA RECOVERING

N.V. Kuznietsova
Institute for Applied System Analysis of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
natalia-kpi@ukr.net

The paper is devoted to new analytical information technologies for clients' preferences prediction. The problem of clients' preferences forecasting is now actual for many commercial systems, companies, banks, insurance companies and e-commerce. Various marketing efforts are used to increase demand and attract new customers. The main idea is to understand the customer needs and preferences and to model their behavior with analytical technologies. Such technologies gives the possibility of the clients' data analysis, customer demand evaluation and prediction of the next purchases. The modern approaches to clients' preferences prediction were analyzed and collaborative filtering methods were chosen. The formulation of the task modelling in terms of clients as subjects and purchases as objects was fulfilled. The method for incomplete and missing data recovering, which was proposed by the author, consists of such stages as sample incompleteness evaluation, analysis if the passes are systemic, analyzing of the passes' causes and effects with using Bayesian network, regression modelling for passes recovering. The method of implicit feedback with the combined method for incomplete and missed data processing were built in the existing modern ERP-system and gives the possibility to receive highest accuracy of clients' preferences prediction.

Keywords: *Implicit feedback, Missed data, Collaborative filtering, Data Recovering, Clients' preferences.*

Introduction

The economic growth of any country, the further development of the economy is accompanied by increasing of the incomes and profits for companies and countries' residents at the same time. There is also a significant development of the customer service sphere, increasing demand of the different products. The client-oriented strategies become the priority for customer service companies. The corporations are included in the intensive competition for clients. The key is the

understanding of customer needs and preferences [1, 2], modeling of consumer behavior, and therefore, most companies and corporations invest heavily in developing their own solutions or purchasing existing ones [3,4]. Such techniques allow them to analyze customer preferences, their pre-orders and develop models that will include interesting for customers products, which can be offered to them for the next order.

1 Clients' preferences analysis

A foreign company specializing in the sale of media products - CDs, DVDs, as well as elite segment products has developed its own Enterprise Resources Planning (ERP) system for collecting statistical information about online store customers, catalogs of goods and knowledge bases on the previous experience. Knowledge are in form of related products recommendations and new acquisitions on the basis of the previous goods purchased in the online store. The company has more than 2 million unique customers and more than 5 million orders statistic. Various marketing efforts such as sending emails are used to increase demand among its users and to attract new customers. The formation of recommendations for related products is carried out by an analyst who selects goods by his own algorithm. Such recommendations are not really precise while people are not able to process such volumes of information. Analyst uses company's ERP system as an automated workplace. The analytics recommendations are stored in the knowledge base of the company, together with information, whether the client used this recommendation.

The company database includes user information, product information, and date of purchase. The database contains missed and lost data. The following simulation tasks are relevant for the company:

- *Forecasting:* sales estimation, server load or server downtime forecasting for providing quick user access to the order directory.
- *Analysis and risk assessment and minimization:* selecting the most promising clients for target e-mailing (the risk of choosing customer who will not buy goods. The losses are calculated as the amount spent on such ineffective mailings).
- *Providing recommendations*: identifying products which can be sold together with the high probability, creating recommendations for preferences.
- *Sequency search*: customer choice analysis while making purchases, forecasting the next possible event.

- *Grouping*: dividing customers or events into clusters of related elements, analyzing and predicting common features.

Modern approaches to analyzing and forecasting the preferences and behaviors of clients within the framework of the advisory systems are ideologically divided into the following types [2,5]:

1. Collaborative filtering methods.

2. Knowledge-based filtering.

3. Methods based on content analysis (Content-based filtering).

4. hybrid methods.

Collaborative filtering is the process of filtering information or samples by sharing multiple technologies, points of review, data sources, and more. Collaborative filtering is usually associated with very large data sets and therefore is appropriate for using in financial systems, Such systems provide financial services, process large amounts of information and combine a large number of financial data sources. In the narrower sense, collaborative filtering is one of the methods for forecast constructing in recommendation systems that uses well-known user group estimates to predict unknown user ratings [3]. The main assumption of collaborative filtration is as follow: those who have equally evaluated any objects in the past tend to give similar assessments of other subjects in the future.

2 Problem statement

The following problem was solved: to develop the information technology for automation the process of recommendation providing on the accompanying products and next products interesting for the clients.

Let there exist a matrix R of size $u \times o$ with the subjects (clients), objects (goods), and some feedback data (previous orders). It is necessary to find a way of transforming it into one matrix with subjects and their profiles (hidden preferences) $P = (p_{tu})_{|T|\times|U|}$ and one matrix with objects and their profiles (hidden preferences that they satisfying) $Q = (q_{to})_{|T|\times|O|}$. The P and Q matrices contain scales that determine how each subject / object relates to each t. The task is to calculate P, Q in such way that their multiply approximates R as closely as possible: $R \approx P \times Q$.

In the process of iterative assignment of random values in the matrices P and Q, using the method of least squares (LS), we must arrive to the same value of the scales that most closely approximate the matrix R.

In the LS algorithm consistently, at each iteration, the following states of the system alternately change:

- P is fixed, then optimizing Q;
- Q is fixed, then P is optimizing.

This operation is continued until approximation to $R \approx P \times Q$ will be reached.

3 Criteria for quality assessing of the customer preferences' prediction

Standard criteria for estimating the quality of forecast such as RMSE or MAE, couldn't be used to assess the accuracy of the solution to the problem of analyzing and predicting customer preferences. It is difficult to estimate if there is a mistake in forecasting model of the clients' preferences or it is the client's decision not to buy this product here and now. Maybe in the future, this customer will buy this product later. It would be advisable to organize the collection of statistical information about the fact of the reference product review, but this is also an indirect characteristic, since the client may not have enough time, but the product is interesting for him, so the recommendation is correct for this client. Such criteria for quality recommendations evaluation [5] were used:

- ✓ **Accuracy:** $\boldsymbol{Precision@k} = \frac{\xi}{k}$, (1)

where ξ – the number of recommended objects with which the subject has an interaction (that is, the number of correctly predicted preferences); k – number of recommendations. This criterion indicates which rate of recommendations corresponds to the preferences of the subject.

- ✓ **Completeness:** $\boldsymbol{Recall@k} = \frac{\xi}{N}$, (2)

where ξ – the number of recommended objects with which the subjects had an interaction, and N – the total number of interactions that was performed by the subjects.

Recall@k – evaluates which fraction of interactions performed by clients corresponds to the predicted interactions, i.e. how many of the forecasted picked goods were interesting to customers.

It is possible to evaluate these equation in cash equivalents, by setting the cost of each interaction and fines for lack of interaction.

4 Implicit feedback and of customers preferences' forecasting

In [6] it was proposed to introduce the following concepts:

$x_{uo} = \begin{cases} 1 & r_{uo} > 0 \\ 0 & r_{uo} = 0 \end{cases}$ – customer preferences:

$\varphi_{uo} = 1 + \alpha r_{uo}$ – level of confidence.

The confidence level is calculated by using the value r_{ui} (feedback, purchases, etc.), which gives more confidence more often, when more often subject interacts with the object. The level of confidence increases due to the linear scaling factor α (which is a "hyperparameter" model). In the confidence level, the constant 1 is always added, indicating $\varphi_{uo} > 0, \forall \alpha r_{uo} \geq 0$ [6].

Then, the mathematical model of the task loss function is formed as:

$$\sum_{u,o} \varphi_{uo}(x_{uo} - p_{tu}^T q_{to})^2 + \lambda(\sum_u \|p_{tu}\|^2 + \sum_o \|q_{to}\|^2) \rightarrow min_{p_{tu}^* q_{to}^*} \quad (3)$$

The component $\lambda(\sum_u \|p_{tu}\|^2 + \sum_o \|q_{to}\|^2)$ is needed to regularize the model in such a way as to prevent retraining. The exact value of the parameter λ depends on the data and is determined by cross-validation.

The loss function contains u × o values. For typical data sets, this value can be several billions. This enormous amount of values impedes most direct methods of optimization, such as stochastic gradient descent, which is widely used for explicit data collection. Therefore, in [6] were suggested an alternative effective process of optimization. If the entities and their profiles or objects and their profiles are fixed, then the loss function becomes quadratic and can be calculated. This statement leads to the use of the method of alternating least squares [6].

4.1 Predicting client preferences

After calculating the preferences profiles of objects and subjects, one can recommend to a particular subject u are K available objects with the highest values of weight $x(u)_i$ – the predicted preferences of the customer u of the product o, namely:

$$\widehat{x(u)_i} = P_i Q^T, \quad (4)$$

where $\widehat{x_{uo}}$ – symbolizes the predicted preferences of the object u of o.

4.2. The task of finding related products

Search for related products can be reformulated as a search for such products o that are similar to preferences and which they satisfy to u customers. Denote the numeric expression as sim_score:

$$\widehat{sim\_score} = QQ_i^T, \qquad (7)$$

5 Preliminary data preparation

A set of inputs is a statistical information about customers who have purchased certain products. The following characteristics are collected: the unique client identifier in the system (Kunden_Id), the categorical variable for clients' gender (Geschlecht_Id - contains gaps), the name (Ort) and index (Plz) of the client's city (incomplete data), the client' birthdate (Geburtsdatum - contains gaps), the unique product identifier (Artikelnummer), the price (Produkt_Preis) and the date of sale (Rechnungsdatum) of the product and the quantity of the product purchased (Anzahl).

That is, there are 4 characteristics with possibly incorrect or missed / lost data. In order to properly handle them, it is proposed to perform a deep analysis of the causes of the gaps' occurrence and to use the combined method for incomplete and lost data recovering, which is proposed by the author. Method consists following steps.

1 step. Estimation of the data incompleteness of the sample for each characteristic by the criterion for estimating the number of passes.

If $I_{j(miss\,ing)} > 20\%$ then, the variable-characteristic is excluded from the simulation and missing values for this characteristic do not make sense to recover.

2 step. Analysis of variables and systematic appearance of missed values

2.1. For a categorical variable assigning missed values to a separate category - filling the spaces with the value:

$V_{категор} :=$ "Missing"

2.2. For all numerical variables with gaps we analyze their appearance (S-systematic):

$$S_{j\,num} = \begin{cases} 1 - \text{for systemiatic passes, where } I_{j(miss\,ing)} \geq 5\% \\ 0 - \text{for non-systematic passes, where } I_{j(miss\,ing)} < 5\% \end{cases}.$$

3 step. Analysis of Causes and Effects.

A Bayesian network is used to establish causal relationships between variables and analyze the consequences of the missing value occurrence. Target (predicted) variable for Bayesian Network – effects.

3.1. To analyze the causes and consequences of the occurrence of passes:

$$C_j = \begin{cases} 1 - random \\ 2 - critical \\ 3 - catastrophical \end{cases}$$

3.2. If for j-th variable $S_{j_{num}} = 0$, $C_j = 1$, then all i-th missed values are replaced as:

$$v_{j_i} = \begin{cases} 0 \\ as \bmod e \end{cases}, \text{ where } V_{j_{num}} = \begin{pmatrix} v_{j1} \\ v_{j2} \\ \\ v_{j4} \end{pmatrix} - \text{i-th value is missed.}$$

3.3. Otherwise, a regression equation is used to predict values.

4 step. Regression modeling.

For linear models the representation in the form of first order autoregression:

$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k)$, $E[\varepsilon(k)] = 0$.

Then, the forecast for one step could be calculated:

$y(k+1) = a_0 + a_1 y(k) + \varepsilon(k+1)$,

If the coefficients a_0, a_1 are known, then the forecast as a conditional mathematical expectation is formed as:

$$\hat{y}(k+1,k) = E_k[y(k+1)] = E_k[y(k+1) \mid y(k), y(k-1), ..., \varepsilon(k), \varepsilon(k-1), ...] =$$

$$= a_0 + a_1 E_k[y(k)] = a_0 + a_1 y(k),$$

For s-steps the forecast is calculated by the function:

$$\hat{y}(k+s,k) = E_S[y(k+s)] = a_0 \left(\sum_{i=0}^{S-1} {a_1}^i \right) + {a_1}^S y(k) = a_0 \sum_{i=0}^{S-1} {a_1}^i + {a_1}^S y(k)$$

.

The sequence of forecasts is a convergent process if the condition $|a_1| < 1$ is fulfilled, that is: $\lim_{s \to \infty} E_k[y(k+s)] = \frac{a_0}{1 - a_1}$, $|a_1| < 1$

Extension of the forecasting function in the process of autoregression AR(p):

$$\hat{y}(k+s,k) = a_0 + \sum_{i=1}^{p} a_i \hat{y}(k+s-i) \ ,$$

where $\hat{y}(k+s-i) = E_k[y(k+s-i)]$.

5 steps. Recovered data application for next simulation.

Applying the proposed method to categorical variable Geschlecht_Id on 2 step, fill it out as "Missing", in this case - 0, and assume that this is "Gender not specified" category. For variables city and zip code (Ort, Plz), fill all gaps as "unknown" and delete all characters except numbers. For the birthdate variable Geburtsdatum, the direct reversals of the gap due to the regression model forecast are not considered appropriate since this characteristic is significant and the gaps can be systematic. While the "age" is usually perceived as the number of full years, it makes sense to do the following. Fill the gaps with the values that correspond to the first day of the first month for the current year. Next, we create a new "Age" variable, which is calculated as the difference between the values of the year in the Geburtsdatum variable and the value of the year for current date. This variable will be an integer and will be greater than or equal to zero. A zero will display a separate case of missed data.

In Fig. 1 the visualization of this data set is presented.

Charts of dependencies between characteristics show that:

- in the customer database is almost equal quantity of women and men, with a minority of women;
- the distribution of orders among cities is close to uniform;
- the distribution of age values for clients has an average value of 61 years old.

A set of attributes for the implementation of collaborative filtering is the following: Kunden_Id, Artikelnummer, Anzahl. It is important to note that quality testing of recommendations will be completed in 2 stages:

- expert of the company sets different user IDs and subjectively analyzes the issuance of recommendations;

- if the first stage is successful, the next stage is performed, namely, the analysis of the accuracy of the model through the Precision@k and MeanAveragePrecision@k

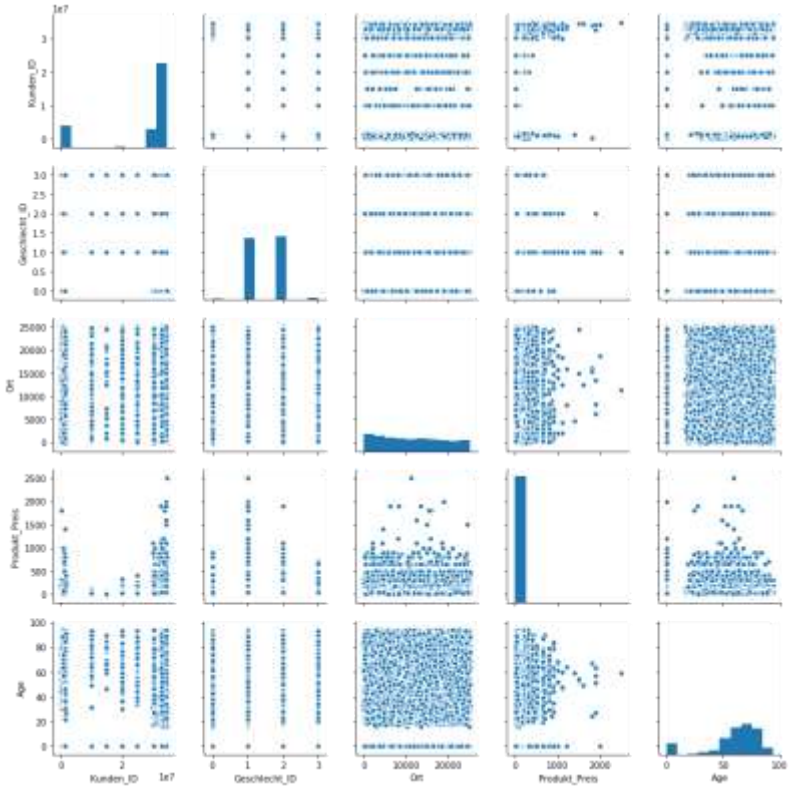


Fig.1 Dependence charts of the characteristics

6 Results of Collaborative Filtration Model Based on Alternating Least Squares (ALS)

Cross-validation, sometimes called cross-check, is a technique of verifying how successfully the statistical analysis by the model is able to work on an independent dataset. Usually, cross-validation is used when the purpose is foresight, and it is important to assess how prognostic model is capable for practice. Cross validation is a way to evaluate the ability of the model to work on a hypothetical test set when it is impossible to obtain such a set explicitly [7].

The model proposed in this paper has hyper parameters:

- Number of iterations - i;
- Number of hidden factors - t;
- The value of the regularization factor is λ;

In all experiments, the quality function will be Precision@k. A "grid" of parameter values is formed as: i - moves from 5 to 50 with a step of 5; t = 50; λ = 0.01.

After performing of 10 iterations, changing the value of the number of iterations, the Precision@k dependency graph is built in dependence from the value of the ALS count iteration indicator and the MeanAveragePrecision@k dependency graph from the ALS count value. By obtaining the first approximation of the optimal iterations number, it is fixed and the optimal value of the regularization coefficient is found. The next "grid" of the parameter values is: i = 10; t = 50; λ - moves from 0.01 to 1 with the step 0.01.

After performing 10 iterations, by changing the value of the regularization factor index, a plot of the dependence of the Precision@k indicator on the value of the indicator λ is built. By changing the value of the indicator, the coefficient of regularization, the dependence of the MAP@k parameter on the value of the indicator λ was constructed. By obtaining the first approximation to the optimal number of iterations and the value of the regularization coefficient, they are fixed and the optimal value of the number of hidden factors is found. The next "net" of the parameter values is formed: i = 10; t - moves from 100 to 1600 in increments of 100; λ = 0.09.

By completing 10 iterations, changing the value of the indicator, the coefficient of the number of hidden factors, we obtained a graph dependence of the Precision@k indicator from the value of t (Fig. 2).

Similarly, by performing 10 iterations, changing the value of the indicator, the coefficient of the number of hidden factors, the plot of the dependence of the indicator MAP@k on the value of the indicator t (Fig. 3) is constructed.

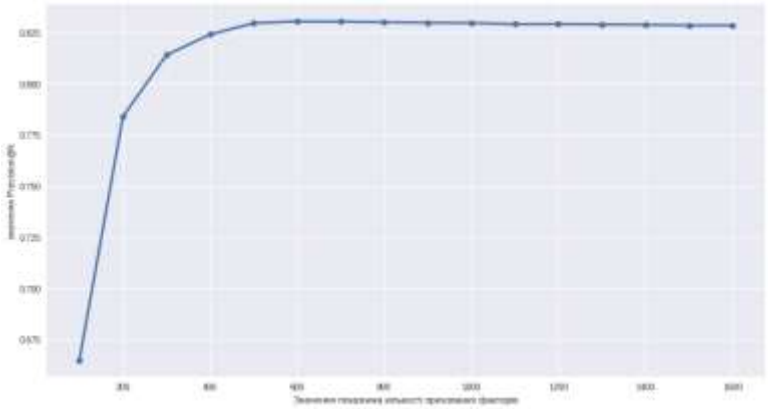


Fig. 2 The dependence of Precision@k on the amount of latent factors

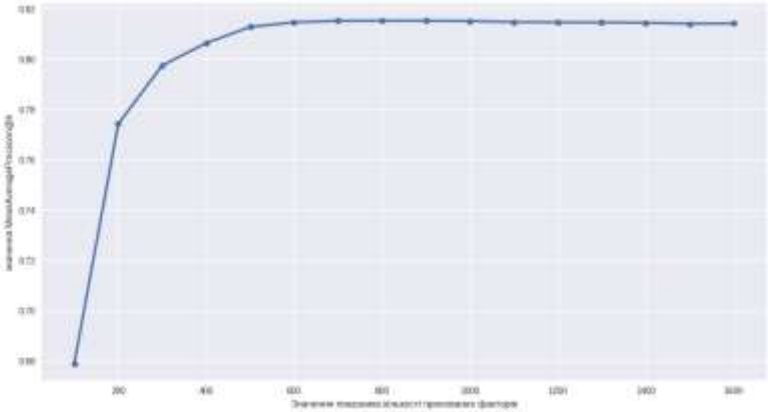


Fig. 3 The dependence of MeanAveragePrecision@k on the amount of latent factors

7 Analysis of the results

Both accuracy functions have a declining character, which is expected, since, with the increase in the number of optimization iterations for each of the model components, the model retraining takes place. For Precision@k, the optimal number of iterations is 15, for the statistic MeanAveragePrecision@k the number of iterations in the value of 10 is

optimal. Therefore, the least of these values was selected. The optimal value for the regularization factor is 0.09 and the values of the received accuracy indicators correlate with each other.

The optimal value of the number of latent factors is 900 (Fig. 2 and 3). The dependency function is increasing to a certain level and after that the level is almost in the same range. This indicates that an optimal number of hidden factors for the given set of data was determined. The number of latent factors is the most important indicator of this system. This is confirmed by the increase in the quality predictions from 67% to 83%. It should be noted that all values of the hyperparameters of the model are relevant only for the set of data that was investigated during the experiment. For new samples, the process of analyzing data on other sets should start again from the beginning according to the algorithm described above.

Conclusions

Ensuring customer loyalty to the company is now a key priority in shaping the relationship between the company and its customers, providing them with high quality products and services which they need. Determining the users' needs and making recommendations when customer is choosing the product it is the main factor in the formation of business models for many companies. Development and using of recommendation systems in the e-commerce market is currently very relevant [9, 10]. The advices of such systems allow companies to use collaborative filtering levers and feature-based recommendations to better serve their customers and increase sales.

Many approaches and methods for recommendation systems constructing have been developed. Most techniques are limited by the fact that they are not able to work on such data as statistics of goods sales, etc. It is necessary to analyze the behavior of users, and for this to determine the key factors of their behavior. Since the entry into force of the "General Data Protection Regulation" Act (GDPR, the European Union), as of May 25, 2016 [8], the personal users' data should not be used without the consent of clients, and therefore such customers should be "forgotten." It turned out to be difficult to work on advisory systems for specific clients. The way of solving this problem is to reclassify clients as 1/0 (old / new client) and extract information from "implicit feedback on binary data", such as, for example, de-personalized statistics on the sale of goods, becoming relevant again. Commercial companies have statistics, therefore, even without the influence of GDPR, the task is relevant.

The methods of searching hidden factors allow solving the problem of analyzing and predicting customer preferences, so are recommended for using in modern business solutions [2, 11].

References

1. Kuznietsova N. V.: Information Technologies for Clients' Database Analysis and Behaviour Forecasting. In: CEUR Workshop. Selected Papers of the XVII International Scientific and Practical Conference on Information Technologies and Security (ITS 2017), 2017, Vol. 2067, pp. 56-62. http://ceur-ws.org/Vol-2067/, last accessed 2018/11/11.
2. Recommendation Systems. Laboratory of Mathematical Logic at PDMI RAS, https://logic.pdmi.ras.ru/~sergey/slides/N16_AIRush.pdf, last accessed 2018/11/11.
3. Data Mining Concepts. Microsoft Documentation Library Homepage, https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/data-mining-concepts?view=sql-server-2017, last accessed 2018/11/11.
4. Data Mining. Base Group Labs Homepage, https://basegroup.ru/community/articles/data-mining, last accessed 2018/11/11.
5. Воронцов К. В. Коллаборативная фильтрация: видеолекции. Школа Анализа Данных Яндекс, https://www.youtube.com/watch?v=kfhqzkcfMqI, last accessed 2018/11/11.
6. Hu Y., Koren Y., Volinsky C.: Collaborative Filtering for Implicit Feedback Datasets. In International Conference on Data Mining 2008, pp. 263-272. Eight IEEE (2008). http://yifanhu.net/PUB/cf.pdf, last accessed 2018/11/11.
7. Cross Validation. LONG/SHORT Blog, http://www.long-short.pro/post/kross-validatsiya-cross-validation-304, last accessed 2018/11/11.
8. EU General Data Protection Regulation Homepage, https://eugdpr.org/, last accessed 2018/11/11.
9. Deshpande M., Karypis G.: Item-based top-N recommendation algorithms, ACM Transactions on Information Systems, vol. 22, pp. 143-177, (2004).
10. Takacs G., Pilaszy I., Nemeth B., Tikk D.: Major Components of the Gravity Recommendation System, SIGKDD Explorations 9 , pp. 80–84, (2007).
11. Вандер П. Дж.: Python для сложных задач: наука о данных и машинное обучение. Спб.: Питер, 576 с. (2018).

USAGE OF EXPERT CLASSIFICATION IN DIAGNOSTIC EXPERT SYSTEMS' KNOWLEDGE BASES CONSTRUCTION

Aleksandr Koval[1], Safwan Al Salaimeh[2], Oleh Andriichuk[3,1]
[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine, avkovalgm@gmail.com
[2]Aqaba University of Technology, Aqaba, Jordan, safwan670@yahoo.com
[3]Institute for Information Recording of National Academy of Sciences of Ukraine, Kyiv, Ukraine, andriichuk@ipri.kiev.ua

Studies of resent years in the field of artificial intelligent led to a number of major achievements. The most significant of these was the development of powerful computer systems, known as systems based on knowledge. They are based on the program, designed to represent and apply actual knowledge from subject domain to problem solving. Diagnostic belongs to weakly structured domains, so that the essential (sometimes only) source of information in this domain is experts.

In this paper the possibility of computerized construction of diagnostic expert systems' knowledge bases is described on the basis of solving the problem of expert classification. The described in the work approach provides a complete build of consistent knowledge bases of diagnostic expert systems.

Keywords: *knowledge base, artificial intelligent, problem solving, expert system, decision-making, possibility.*

Introduction

Studies of resent years in the field of artificial intelligent led to a number of major achievements. The most significant of these was the development of powerful computer systems, known as systems based on knowledge. They are based on the program, designed to represent and apply actual knowledge from subject domain to problem solving. A general overview of intelligent computer systems is translated into [3]. Example of development in recent years in [2] the possibility of using expert system (ES) is considered for teaching students using a computer, an expert rating system in commercial banks is considered. In [4] the principles of constructing expert systems of real time and for the problem of diagnostics are considered and management and environmental monitoring.

Diagnostic decisions are influenced by numerous quantitative and qualitative criteria, factors, and parameters. It is problematic to provide a formal mathematical (analytic) description of these factors. This leads to the fact that diagnostic belongs to weakly structured domains [5, 6].

1 Features of Weakly Structured Subject Domains

As shown in Fig. 1, there are following properties of weakly structured subject domains: absence of functioning goal which could be formalized, absence of optimality criterion, uniqueness, dynamics, incomplete description, presence of subjective human factor, impossibility of analytical model building, lack of benchmarks, high dimensionalities.

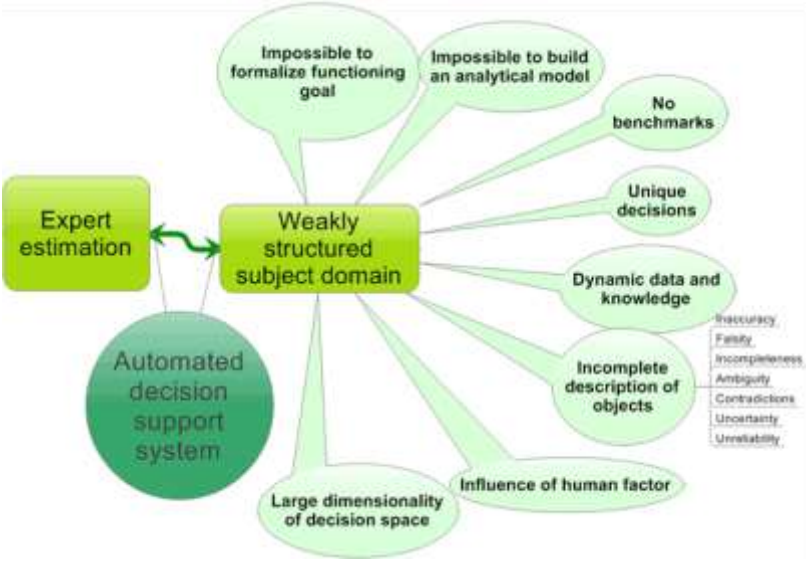


Fig. 1. Features of weakly structured subject domains

Let's consider in more details the characteristics of mentioned above weakly structured subject domains. Objects in such domains are unique. In fact, management systems for these areas are created once, in order to solve real problems; the transfer of such models to other objects is costly or simply impossible.

In systems that are not created by a human (such as biological), there is no goal of functioning that can be formalized. The purpose of the functioning of such systems is their efficiency in general, the support of some parameters in the specified limits, but it is usually impossible to formalize such a goal in the form of a certain criterion. For instance, in the biological system, all factors affecting its functioning are so numerous and the connections between them are so complex and not obvious that it is impossible to set a certain function to describe the purpose of its functioning.

Due to the lack of a functioning goal that is subject to formalization, it is impossible to construct a function whose optimization would provide the best mode of operation of the object. Objective optimization function does not exist, you can specify only some of the factors that can be optimized. However, it is impossible to optimize each of these factors separately, since they are closely interconnected, and their connections will be disturbed during the process of optimization. This can lead to a violation of the process that regulates or maintains the system in a stable condition relative to the changing environment in which the system operates, and may lead to catastrophic events and irreversible changes in the system.

Since in the weakly structured subject domains there is no goal of functioning that can be formalized, and it is impossible to construct a function whose optimization will provide the best mode of operation of the object, then it is impossible to construct an analytical model of this subject domain.

Dynamism is due to the fact that the structure and functioning of the object change over time, that is, the object evolves. Management of such systems should be adaptive, able to change when the object changes.

The incompleteness of the description is due to inaccuracy, incompleteness, falsity, ambiguity, contradictoriness, uncertainty, and unreliability of the data describing the object.

Characteristics of objects is problematic to describe quantitatively, therefore in weakly structured subject domains it is inappropriate to speak of the existence of benchmarks of these characteristics.

The large dimension of the decision space is due to the large number and heterogeneity of the criteria that characterize the subject domain.

The objects of management can be people who have free will. It is often impossible to predict human behavior as an object of control or a component of a system. A person operates in the system, taking into

account his personal goals and interests. Therefore, when modeling the object of human behavior management is difficult to consider.

Described above properties of the weakly structured subject domains lead to the fact that the essential (sometimes only) source of information in them is experts.

2 Problem Statement

A fairly compute analysis of the problems organization – related receipt and the interpretation of expert information when creating an expert system is contained in the work [7].

Wide class of task, for which expert systems are development, design diagnostic tasks. Diagnostic expert systems are essentially designed to solve the classification problem: each object (situation). From the subject area a sings his diagnosis. Properties, course of action large role in the experts work of their knowledge is plays with the manual of the organization of expert inquiry. Success in creating an information system became possible due to the fact that a certain expert apparatus was developed, allowing to largely overcome the above problems. The essence of this approach is use of data on the structure of the problem to be solved, about the relaxation between its other elements. Such information makes it possible to draw certain conclusions on the possible limits changes of a number of estimates and thus reduces the number of questions to the expert, control the consistency of information. Identify and eliminate appearing in it mistakes [8, 9].

Diagnosis of the object on the basis of its description is possible only if the different values of the signs have different degrees of characteristics and for diagnosed properties. In connection with these, when structuring the task of expert classification, hypnosis is used about the varying degrees of the characteristics of the individual values of each characteristics and for each property (or what is the same class). It is also assumed that, for each feature, the expert may order its value by their characters and for the corresponding class and this order doesn't depend on the value of other characteristics. The experiment of solving the problem of expert classification suggests that the assumption of ordering the meaning of the attribute is also valid for many practical problems, and accordingly, the considered formulation covers a wide class of the problem of expert classification, the possibility of obtaining valuable information about the objects membership class makes it possible to construct a rational procedure for expert demand in order to minimize the

number appeal to him. In addition, detailed information allows you to identify possible errors in the expert's answers [10, 11].

3 The solution of Problems

The formulation of the classification problems is as follows:

Given a set $P = \{P_1, P_2, \dots, P_i\}$ independent properties which can have the object of research; M signs, characterizing from different sides the objects of research; set $Q_m = \{q_{m1}, q_{m2}, \dots, q_{mn}\}$ possible *m*-th value of the *m*-th characteristics and n_m – the number of these values $A = Q_1 \times Q_2 \times \dots \times Q_m$ - set of all hypothetical possible states of the object of investigation, in this state $a_i \in A$ is characterized by a vector $a_i = (a_{i1}, a_{i2}, \dots, a_{im})$where $a_{im} \in Q_m$, $m = \overline{i, M}$. Formally hypotheses a specifically can be described as follows, ordering characteristics values Q_m by their character for the property p_i allows to add to Q_m transitive and antifriction binary relations (linear order) $(q_{ms}, q_{mi}) \in r_m$, if q_{ms} more characteristics for this property than q_{mt}. on this basis of these relation it is possible to construct a binary relation of dominance by characteristic for each property on the set of the state on the object of investigation which are described as follows:

$R = \{(a_s, a_i) \in A \times A |\ m = \overline{i, m},\ (a_{is}, a_{im}) \notin r_m\}$ and $m \in M: 1 \leq \vartheta \leq m$, such that a, $a_{s\vartheta}, a_{i\vartheta}, a_{t\vartheta} \in r_\vartheta$ have a reflexive transitive relation of strict dominance.

Required based on expert knowledge for each state from A identify the presence of relevant and properties from the set P and thus build a classification of the set $A = \bigcup_{i=0}^{k} k_i$ such that, the state $a_i \in A$belongs to the class K_i , if the object in this state has the experts opinion and properties P_i. To class K_o, in this case, there are such states in which the object doesn't possess, in the experts opinion, not one consideration. It is customary to call this problem the main task of classification [12,13,14].

The procedure for constructing the classification is described as follows: we put in correspondence to each state $a_i \in A$ to sets C_i^+ – set of the numbers of classes belong to the state a_i and C_i^- – set of class numbers, to which the state a_i cannot belong. Let $K = \{0, 1, \dots, L\}$ – multiple numbers of generated classes. Status will be considered classified, if $C_i^+ \cap C_i^- = \emptyset$ and $C_i^+ \cup C_i^- = K$, denote by $A^o \in A$ a subset of all classified state. Before the beginning of the expert survey $\forall_{a_i} \in A$ suppose $C_i^+ = C_i^- = A^o = 0$. The procedure for an expert survey and when $A^o = A$. In the course of the survey the next required expert advisor

is selected $a_i \in A \backslash A^o$. The expert makes conclusion on this condition in the form of the list of classes belonging (object properties, is in this state). Thus for this state a_i the set is defined explicitly C_i^+ and an implicit set $C_i^+ = K \backslash C_i^-$. After this condition a_i is classified and $A^o = A^o \cup a_i$. Information received regarding the state a_i, allows reducing uncertainly for a number of other states [15 - 20].

Along with the main task of expert classification on solving practical problem, there may be a need to clarify the severity of the diagnosable properties, which can be sensible ordered. With this definition of groups the state of the objects, in which he has the same degree of severity and diagnosable properties, means the spitting of the state of an object into a set of defined classes. It is customary to call this task the expert classification of the problem of order to classification.

The problem of the order of expert classification for a certain property can be defined as follows: there are many $P = (P_1, P_2, \dots, P_n)$ independent of properties, which may have the object of study. For some properties of P determine the set of its values $\{P_1, P_2, \dots, P_n\}$. on this set the ration of linear order is determined $\check{R}$such that $(P_i, P_j) \in \tilde{R}$, if $i < j$ (i.e. value set are ordered from larger than the degree of expression of this property).

Let $Q = \{Q_1, Q_2, \dots, Q_m\}$ – set of symptom, describing the state of the object of study. On the scale of each feature Q_m $(m = \overline{1, M})$. A certain ratio of linear order $(q_{ms}, q_{mt}) \in r_m$, if q_{ms} more characteristic for this property, than q_{mt}, $A = Q_1 \times Q_2 \times \dots \times Q_m$ the set of all possible states of the object of investigation. On this set a certain relation of strict domination *R* (similar to the considered above).

Required based on expert knowledge. Determine the severity of this device for $\forall a_i \in A$ and thus to construct a partition of the set $A = \bigcup_{n-1}^{N} Y_n (Y_i \cap Y_j = \emptyset,\ i \neq j,\ i, j = \overline{1, n})$ such that $a_i \in A$ belong to the class Y_n $(1 \leq n \leq N)$ if the objects in this state have this property of degree P_n.

The procedure for experts can be described as follows: let G_i – a set of class numbers Y_B, which correspond to the state $a_i \in A$. Before the survey for $\forall_{a_i}$, $G_i = \overline{1.N}$, the classification can be considered complete, when $\forall_{a_i} |G_i| = 1$. Let the expert determined, what the condition is $a_i \in A$ corresponds to the value $P_n (1 \leq n \leq N)$ by the degree of expression of the property under consideration, i.e. $a_i \in Y_B$. Consequently, in this case, the state described by a set of characteristics is no les

----------s characteristics for this property, it cannot have a lesser degree of its expression that is, if $a_i \in A$ and $(a_i, a_j) \in R$, then $a_j \overline{\in} Y_n$, $k > n$. Similar condition, described by a set of characteristics value of the characteristics most characteristics for this properties, no more characteristics of this property can have a greater degree of its expression, that is, if $a_i \in A$ and $(a_i,\ a_j) \in R$, then $a_j \overline{\in} Y_n$, $k < n$.

There are various ways to determine the in formativeness of the object to the expert. It is also possible to estimate the informative state is the number of indirectly classified states based on the characteristics relationship. For each state, you can find this number with every possible expert answer and calculate in the middle value or the minimum of them.

Using these indicators it is possible to compare all unknown states of an object by informative their presentation to the expert and choose the most informative. When presentation the most informative status to the expert, we will receive on average the largest amount of information with any expert answers.

Whenever any procedure for interviewing experts should consider the possibility of errors in the answer. These errors can be detected by inattentive, expert fatigue, and also the complexity of the problem being solved. Since the knowledge base must be non-selective, information analysis needed, obtained from an expert, controversy detection. The possibility of indirectly defining the classes of accessories of the state allows you to check the consistency to expert assessments. If there is discrepancy in the indirect and direct evaluation of the status this indicates that there are errors in his answer. It is necessary to present contradictory answers of experts for their comprehension and choosing the right way to assess the series condition.

Possible two strategies for removing the contradiction when building a knowledge base. One of this to continuously compare information, receive from the expert, from the received early and check for inconsistency. If there is contradiction between the last expert answer and the previous information, this contradiction is presented to the expert for analysis and choice of the contradiction of politics. Another strategy is to obtain from the expert or a part, either all the necessary information, and only then the implementation and in it of the search for contradictions and their removal.

In this way. The procedure developed by the expert survey should, on the hand, minimize the work of the expert and on the other hand, allow him to analyze the information received from him, from the point of view

of its consistency. If the traditional way of building a knowledge base for diagnostics systems, an expert has to solve the problem of synthesizing of their knowledge, then the proposed method corresponds to its usual task of analysis specific situations. So, he unconsciously uses many of his skills and techniques, which is difficult for him to formulate in an explicit form.

The methods for solving the problem of expert classification have natural limitations, due to their dimensionality. In problems of large dimension (dozens of the signs with a large number of possible boundaries and a large number of diagnosed properties) the laboriousness of the experts work also sharply increases and the computational complexity of the procedure increases the solution of the problem of expert classification, which makes it difficult to use them directly. In practical problems of classification, it is often possible to distinguish separate groups of attributes according to their semantic content in such a way that the characteristics to one group reflect on of the sides of the object under consideration. In these cases one of the possible ways of constructing the coverage of the initial state set is to decompose the original problem into subtasks less than the size, the used classification procedure will consist of several stages.

Let all sets of attributes Q broken into T groups: $S_1, S_2, \dots, S_T$; $U_{S_i} = Q, S_i \cap S_j = \emptyset$. Each group of signs to a certain extent characteristics the presence of the object to be diagnosed to properties. Therefore, according to the values of the *i*-th group of characteristics ($i = \overline{1,T}$) the expert can make a preliminary conclusion about the possible classes of states, the characteristics described by these values.

Each group of characteristics gives incomplete, only a partial description of the state of the object of investigation. Accordingly, the evaluation of the expert, it is based only on the part of the object description, can has a probabilistic character and reflects the varying degree of the experts confidence in the availability or the absence of the object in the given state of the diagnosed properties. Consequently, at the first stage of solving the problem, the expert decides its main, and order any classification tasks within each group of characteristics, which allows to build an orderly classification, and for each properties for each group of each characteristics. This give opportunity for each state $a_i = (a_{i1}, a_{i2,}, \dots, a_{iM})$ and for each property P_i define a vector $(T_i, T_2, \dots, T_S)$, where T_S – degree of suspicion of *i*-th property in the state a_i by *j*-th group of attributes.

The next stage of the solution consists in construction generalized classification for each property based on the other of their classification by separate groups of characteristics. In this case, the signs of the second level are also the degree of the experts confidence in the availability of the *i*-th property for each of the *T* group of initial signs. It is necessary for each property P_i $(i = \overline{1,l})$ on the basis of the value of the attributes of second level, construct an order classification, determine whether or the absence of each of the property under consideration and its degree of serenity.

If the expert cannot make a diagnostic conclusion for a certain group of signs. Not having information about the result, received for another group of characteristics, the survey is conducted as follows: if the diagnosis is on *i*-th group, the expert must build *L* order of classification for *i*-th group of characteristics with the addition of the result for property p_i by *j*-th group of characteristics, number of hierarchy levels (as the decomposition levels) can vary in different tasks. It is determined the classification problem.

Conclusion

In this paper the possibility of computerized construction of diagnostic expert systems' knowledge bases is described on the basis of solving the problem of expert classification. The described in the work approach provides a complete build of consistent knowledge bases of diagnostic expert systems.

References

1. Safwan Al Salaimeh The Optimal Management of Information Servicing Logistics System / Institute Mathematics and Computer Science Journal – India, 2003. – pp. 75-80.
2. Safwan Al Salaimeh Information Technologies of Distributed Applications Design / Institute Mathematics and Computer Science Journal – India, 2003. – pp. 99-103.
3. Safwan Al Salaimeh, ZaferMakadmeh Multi-Criteria Synthesis of Logistics Systems Through the Hierarchy Analysis / Journal of System Sciences // Poland University of Technology, Poland. – pp. 107- 115,
4. Safwan Al Salaimeh, Khaled Batiha Business Process Simulation with Algebra Event Regular Expression / Information Technology Journal – Volume 5, Number 3 – Pakistan 2006 – pp. 583-589.

5. Taran T.A., Zubov D.A. Iskusstvennyy intelekt. Teoriya i prilozheniya (Artificial intelligence. Theory and applications) // Lugansk: V. Dal VNU, 2006, - 239 p. (in Russian).
6. Glybovets M.M., Oletsikiy O.V. Shtuchnyy intelekt (Artificial intelligence). – K.: "KM Akademiya" pulblishers, 2002 – 366 p. (in Ukrainian).
7. Khaled Batiha, Safwan Al Salaimeh E-Learning / Leonardo Electronic Journal of Practices and Technology – Romania, 2006. – pp. 1-4.
8. Khaled Batiha, Safwan Al Salaimeh, Khaldoun Al Besoul Digital Art and Design / Leonardo Journal of Science – Romania 2006. – pp. 1-8.
9. Khaldoun Al Besoul, Safwan Al Salaimeh The Structure of logistics organizational technological system / Journal information society – Vol.4, Num. 7 – Romania, 2007, pp. 126-129.
10. Howard Bandy Modeling Trading System Performance / Blue Owl Press, Inc., 2011. – 384 p.
11. Yan Houmin, Yin George, Zhang Qing Stochastic Processes, Optimization, and Control Theory: Applications in Financial Engineering, Queuing Networks, and Manufacturing Systems – Springer, USA, 2006. – 360 p.
12. Process dynamics modeling, analysis, and simulation by B. Wayne Bequette 1st Edition / Prentice-Hall PTR, Upper Saddle River, New Jersey 07458, 1998. – xviii + 621 pages.
13. Otto Bretscher Linear Algebra with Applications, 4th Edition / Pearson, 2008. – 504 p.
14. Software process modeling by Silvia T. Acuna, Natalia Juristo / Springer US, 2005. – xxiv + 208 pages.
15. Hossein Bidgoli Modern Information Systems for Managers / Emerald Publishing Limited, 1997. – 438 p.
16. Howard Bandy Modeling Trading System Performance / Blue Owl Press, Incorporated; 1 edition, 2011. – 400 p.
17. Salim Istyaq, Safwan Al Salaimeh, Amjad Miqdadi Decomposition Algorithm of the Model of Electronics Systems for Modeling in Conditions of Distributed Resource / International Journal of Emerging Technology and Advanced Engineering – Volume 8, Issue 3, March 2018. – pp. 29-31.
18. Daniel H. Pink When: The Scientific Secrets of Perfect Timing / Riverhead Books, 2018. – 272 p.
19. Dan Ariely Predictably Irrational: The Hidden Forces That Shape Our Decisions – HarperCollins Publishers, 2008. – 280 p.
20. Marjory Harris The Personal Power Roadmap: The Ultimate 7 Step System to Effectively Solve Problems, Make Decisions, and Reach Your Goals, Paperback – CreateSpace Independent Publishing Platform, 2016. – 168 p.

CHINESE LEGAL INFORMATION AUTOMATIC SUMMARIZATION

Dmytro Lande[1], Zijiang Yang[2], Shiwei Zhu[2], Jianping Guo[2], Moji Wei[2]

[1]Institute for information recording of NAS of Ukraine, Kyiv, Ukraine

[2]Information research institute of Shandong Academy of Science, Jinan, Shandong, China

Article is devoted to a method for automatic text summarization of the legal information provided in Chinese. The structure of the summary and the model of its formation is considered. Two approaches are offered. First one is determination of weight of separate hieroglyphs instead of words in the texts of documents and summaries for sentences importance level determination process. Second approach is to consider a model of document as a network of sentences for detection of the most important sentences by parameters of this network. Various methods of automatic text summarization are performed and tested. A cosine measure and Jensen-Shannon's divergence are applied as two estimates of summaries quality without participation of experts. Compared to other summarizing methods, given one on the basis of the offered network model of the document was the best by criteria of a cosine measure and Jensen-Shannon's distances for summaries which volume exceeds 2 sentences. The offered approach, with minimal modifications, can be applied to texts on any subject of scientific, technical or news information.

Keywords: *Automatic text summarization, Legal information, Chinese language, Cosine measure, Jensen-Shannon's divergence*

Introduction

Processing of natural languages practically began with statement of problems of the artificial translation and automatic text summarization. The first fundamental works on automatic text summarization appeared in the middle of the last century [15].

The task is connected with the solution of the most important problem – reduction of the volumes of information consumed by the person, fight against information noise. This task is very relevant today due to the

constant growth of the information space. Automatic text summarization is known to all users of network search engines - in response to the request they receive not only the title of documents, but also their short automatically created descriptions (snippets). Mobile users want to see a brief description of the articles before they go on to read more. Persons taking important management decisions should acquainted with thousands of documents a day, deliberately dismissing information noise.

Now there are hundreds of industrial systems of automatic text summarization, for example, such packages as Office Word AutoSummarize, Mac OS X Summarize, IBM Tivoli Monitoring Summarization and Pruning Agent, Oracle Text, plug-ins for browsers Chrome, Mozilla.

Numerous approaches to automatic text summarization are known, recently, neural network technologies, deep learning are applied more and more widely. There are also numerous linguistic approaches associated with automatic parsing of sentences submitted in different languages. Traditional type of systems of automatic text summarization – extractive (quasisummarizing) at which the summary consists from of the separate, sometimes weakly connected among themselves sentences of the initial document. He is succeeded by abstractive type of text summarizing at which the systems close to the systems of artificial intelligence in a short form retell contents of the initial document "by the own words".

However it should be noted that today still practically all industrial systems of automatic text summarization belong to extractive systems.

It would seem, the subject of automatic text summarization of texts is already rather studied, the main results are received. However and in this article it is about creation of system of automatic text summarization.

There are several reasons for development of the new automatic text summarization system of automatic text summarization. First the problem of automatic text summarization of legal information is solved. And it is texts which can't fully be considered free, unstructured. There is a structure of separate types of documents and use of the best universal systems of summarizing doesn't yield satisfactory results. Secondly, the authors deal with the texts of documents presented in Chinese, which significantly narrows the range of possible ready-to-use systems. For processing Chinese texts, as a rule, segmentation of words is required – in the Chinese language words are often not separated by separators.

Thirdly, the program capable in corporate system to process big data flows with an acceptable productivity and quality, built in the existing system of document flow has to be developed.

Besides, retelling of documents in this case is unacceptable. Any "imaginations", liberties of retelling by the computer of legal acts it is inadmissible. Exit one – to develop some hybrid algorithm and, respectively, the program of extractive type capable to consider features of legal acts of the People's Republic of China. At the same time the program has to be capable to process separate documents which unite to big documentary arrays. This program has to be capable to allocate obviously set objects in the parts of documents marked with semantic markers, to reveal the most important parts of documents (including by statistical criteria), to form networks of sentences and to remove the necessary volume of target information in the summary.

The offered approach

In addressing the problem, two approaches were proposed that could be considered new in this area. To solve the problem of determining the level of importance of individual parts of documents (in our case, sentences) it was suggested to move to the definition of weight values of separate hieroglyphs, not words in the text of documents and summaries. It was also suggested that the document model should be considered as a network of sentences to identify the most important sentences for the parameters of the network. The weight of the links of the two sentences in this network is determined by the weight of the common hieroglyphs included in them.

Within the traditional statistical approach to the processing of natural languages, the weight of sentences is usually calculated on the basis of the estimated weights of lexical units (words, phrases) included in these sentences [20], [16], [2], [4]. In this work it is proposed as such elements for the Chinese language to use separate hieroglyphs.

The transition from the words considered in the classical model to hieroglyphs allows avoiding the relatively complex procedure of words segmentation the text, which is inevitable with all other meaningful methods of Chinese texts automatic analysis. Of course, this approach is not applicable to European languages, where the number of different letters does not exceed several dozens. However, for the purpose of automatic text summarization of Chinese texts, the proposed approach provides acceptable results, which will be shown below.

It is known, that in the Chinese language there are more than 40 thousand hieroglyphs, therefore each of them (though not always, fully reflecting a semantic unit) it is possible to attribute a weight value calculated on the known formulas, for example, $TF \cdot IDF$ [17].

$TF \cdot IDF$ (*TF* — term frequency, *IDF* — inverse document frequency) is a statistical measure used to evaluate the importance of a word (in this case, not a word, but a hieroglyph) in the context of a document that is part of an array of documents. The weight of some hieroglyph is proportional to the number of its use in the document, and is inversely proportional to the frequency of occurrence of this character in all documents of array.

Thus, the measure $TF \cdot IDF$ depends on the word t (hieroglyph), the document d, the whole array of documents *D*, and is a product of two factors:

$$TF \cdot IDF(t, d, D) = tf(t, d) \times idf(t, D).$$

Here the expression $tf(t, d)$ is the ratio of the number of occurrences of some hieroglyph to the total number of characters in the document (to the length of the document, actually). Thus, the frequency of the hieroglyph within a single document is estimated.

The second factor, $idf(t, D)$ (inverse document frequency — the reverse frequency of the document) is the inversion of the frequency with which some hieroglyph occurs in the documents of array D.

IDF accounting allows you to reduce the weight of hieroglyphs that occur very often. There is only one IDF value for each t within the entire array of the documents *D*:

$$idf(t, D) = \log \frac{|D|}{\left|\{d \in D \mid t \in D\}\right|}.$$

In addition, unlike classical approaches to the definition of weight values of sentences, a new, network model is proposed. Under this model, a non-directional network is considered, with nodes appearing as separate sentences in the document, between which the links are established if they have common hieroglyphs. The weight of the relationship between the two sentences is defined as the sum of the weights common to these sentences. For this network, the weight of each offer is calculated as the sum of the weights of the links of all links that emanate from the node.. Naturally, the weight of the proposals is then normalized, since long sentences without this procedure will on average have a deliberately greater weight. Practice has shown that a good normalization is the division by the logarithm of the length of the corresponding sentence

Automatic text summarization of legal information

Procedures of automatic text summarization of an extractive class are based on determination of weight values (importance degree) of separate sentences which, in turn, depend on scales of words. In work as weight word meanings the classical criterion $TF \cdot IDF$ was used though it is not only are possible for the solution of a problem of summarizing approach [11]. Traditionally for definition of weight word meanings two known algorithms were used – in the first case the weight of the sentence was considered as the sum of weights, rated on length of this sentence, of the words entering it, and in the second case was used, so-called, an symmetric summarizing algorithm [19]. In this case the weight of the sentence r was defined as the sum of weights of its links with the previous and subsequent sentence.

In addition, this paper proposes a network algorithm, which, unlike the second case, calculates the relationship not only between adjacent sentences, but also between all the sentences in the text. This approach, of course, is computationally more complex than the first two, but, as practice has shown, leads to better results. At the same time, the complexity of the algorithm, in the case of the considered approach of texts summarization in Chinese, is compensated by the fact that instead of words (segmentation of which in this case is not required) are considered only separate characters.

So let's present the basic steps of three considered algorithms of definition of weight values of offers:

Step 1. For each hieroglyph t_i the value $DF = df\left(t_i, D\right)$ is calculated as the number of documents d_j from the documentary array D that contain this hieroglyph, that is

$$DF := \left|\left\{d_j \in D : t_i \in d_j\right\}\right|.$$

Step 2. For each hieroglyph t_i and document d the value as frequencies $TF = tf\left(t_i, d\right)$ of emergence of this term in the document is calculated:

$$TF := \frac{\#\left\{t_i \in d\right\}}{\left|d\right|}.$$

Then hieroglyph weight is calculated

$$w_i = TF \cdot IDF = TF \cdot \log \frac{\left|D\right|}{DF}.$$

Step 3. Segmentation of sentences, i.e. the text of the document is divided into separate sentences p_i and then the definition of their weight values wp_i. Let's introduce notation: let the sentence p_i of the set of sentences P ($p_i \in P$) consists of hieroglyphs $t_{i,k}$ with weight $w_{i,k}$. Let's write in a brief form the essence of three different algorithms.

Step 4 a) Algorithm of the sum of hieroglyphs weights ($\sum tf \cdot idf$):

$$wp_i = \frac{1}{|p_i|}\sum_{k=1}^{|p_i|} w_{i,k}.$$

Step 4 b) Symmetrical algorithm for calculating the power of connection of the sentence p_i with the nearest sentences (Nearest):

$$wp_i = \frac{1}{\log|p_i|}\sum_{k=1}^{|T|}\left(w_{i,k}w_{i-1,k} + w_{i,k}w_{i+1,k}\right),$$

where T is the general composition of the hieroglyphs of the array. If the character is not present in the document, its weight in it is equal to zero.

Step 4 c) Network algorithm of calculation of force of link of the sentence (Network):

$$wp_i = \frac{1}{\log|p_i|}\sum_{\substack{j=1,\\ j\neq i}}^{|P|}\sum_{k=1}^{|T|} w_{ik}w_{jk} \ .$$

Step 5. The weight of the sentence is corrected depending on its location in the document. Weight values of initial and last sentences of the document artificially increase.

It should be noted that the specifics of legal information, requirements to the structure and volume of the summary, allowed to use the above-mentioned universal approaches to the solution of a private special task.

The structure and volume of the summary of the legal document (examples of such documents can be found on the website http://www.gov.cn/in the section /zhengce) are put forward requirements that have found their programmatic implementation:

1. Summary start with the title of the document, given almost without changes.
2. The summary notes the type of document (announcement "通告", report "报告", results of work "工作成果", provisions "政策" etc.).
3. If the document indicates its purpose ("目的", "奖补目的", "调整目的", "普查的目的和意义", etc.), it is also reflected in the summary.
4. If the first or second sentence of the document identifies the subjects

of appointment of documents (which is also visible by special markers), such a proposal is also included in the summary.

5. If in the title of the document or in the designation of its purpose explicitly there are objects from the number of the previously known (included in the base objects table), these objects should be highlighted in the summary.
6. EIf the document belongs to the type not subject to further processing (awards "表彰', announcements of bids "招标', letters "函' etc.), the summary is considered prepared..
7. All sentences containing the objects selected from the title and purpose are selected from the text of the document. If such proposals are less than the required number (given in advance or calculated on the basis of the volume of the document), they are presented in the summary in the same sequence as in the primary document. The summary is considered prepared.
8. If the sentences are more than the required number, they are weighed according to the above algorithm (based on the results of testing the network algorithm is selected). After that the sentences ranked by weight and are presented in the summary in the same sequence as in the primary document. The summary is considered prepared.

According to the submitted requirements the program of automatic text summarization of the legal information provided in Chinese was developed.

Adjacent tasks

Automatic text summarization of texts is one of important problems of technologies of the deep analysis of texts which includes some more directions, such as extraction of entities (Information extraction), creation of networks of words (Language Networks) reflecting features of subject domains, a clustering (Cluster Analysis).

The algorithm offered for summarizing leans on some set of in advance prepared words reflecting the main objects presented in legal documents (for example, "人口" – the population, "产业" – the industry, "儿童" – children, etc.).

At the same time, if you apply the algorithm of words segmentation, and then rank them, it is easy to identify the most common "extensions" of starting objects, for example, the concept of "organization" (组织) to expand to the concept of "international organization" (国际组织), "public organization" (社会组织), and the concept of "defense" (事业) to the concept of "people's air defense" (人民防空事业). As a result, the documents of the

array of legal information have been put in line with the basic concepts that can act as "keywords", descriptors, basis for the construction of domain models (Subject domain).

As one of the types of domain models can be considered a words network, the nodes of which correspond to separate concepts. There were proposed and implemented such simple rules of building this network, i.e. rules of communication between nodes:

1. All objects from the base, pre-prepared list, included in one document are linked by links.
2. If two objects are in *N* different documents, the force of link between them equals *N*.
3. Concepts that are extensions of concepts from the starter kit are linked with the corresponding basic concepts.

With the help of the program Gephi (http://gephi.org) [3] the built network has been visualized (Figure 1) and received such parameters of the built network: number of nodes: 3364 (number of objects from the starting set – 220); - number of links: 10167; network density: 0.001; number of connected components: 6; average path length: 3,013; average clustering factor: 0859.

The topology features of the built network include a very large average clustering factor. This is due to the ode of a large number of concepts related only to the natal of their concept (the absence of other neighbors), and on the other hand the strong cohesion of objects from the start list. The small average length of the path indicates that the network is a Small World [8].

With the help of the program Gephi also received lists of the most important nodes in accordance with the criterion of PageRank and the greatest hubs by the criterion of HITS [12] (Figure 2).

The general view of network of words given on the Figure 1 clearly demonstrates a further possibility of a clustering of network, the choice of subsets – clusters from words (concepts). This procedure allows to allocate thematic subsets within the considered subject domain.

Methods for results evaluation

To evaluate the results, two assessments of the quality of the summary are applied without experts – the cosine measure and the divergence of Jensen-Shannon (Jensen – Shannon), the justification of which is substantiated in the work [14].

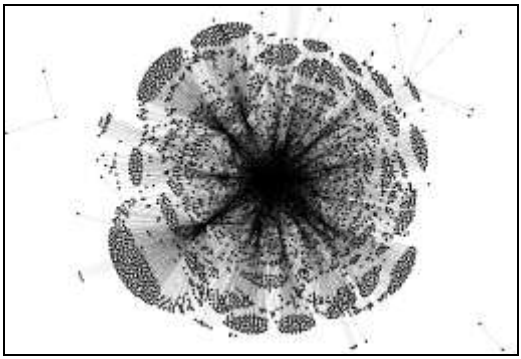


Figure 1. A subject domain words network

Label	PageRank
水利	0.001548
"十一五"	0.00154
扶贫	0.00149
毕业生	0.001408
银行	0.001405
上海市财政	0.001383
邮政	0.001374
林业	0.001352
信息传输	0.001349
城市规划	0.001337
文物	0.001323
医生	0.001315
省财政	0.001242
技术服务	0.001215
山东省财政	0.001194
农民工	0.001126
景域	0.00111
农田	0.001097
电信	0.001072
经济特区	0.001055
科技创新中心	0.001045
试验区	0.001
北京市财政	0.000989
食品药品	0.000964
电影	0.000949
房地产	0.000897
矿产	0.00088
供销	0.000877

PageRank

Label	Hub
残疾人	0.04637
租赁	0.044748
科技创新中心	0.043809
农民工	0.043738
统计	0.04244
电信	0.04112
省财政	0.040688
经济特区	0.039118
山东省财政	0.039081
作业	0.038172
食品药品	0.036646
北京市财政	0.036476
水利	0.036002
试验区	0.035958
电影	0.031812
人工智能	0.031356
娱乐	0.03121
邮政	0.028793
物流业	0.027323
海关	0.026766
社会信用体系建设	0.026739
餐饮	0.02619
深圳市市场	0.02569
干部	0.025645
公共管理	0.025336
金融业	0.025252
食品药品监管	0.025083
经济体制	0.025021

HITS

Figure 2. Most rating words in the criteria of PageRank and HITS

Let us explain the possibilities of using these approaches. The document's d hieroglyphic dictionary is supposed to consist of N elements $\{t_1, t_2, ..., t_N\}$. Each hieroglyph corresponds to its weight, calculated according to the rule $TF \cdot IDF$. An array of these weights can be represented as a vector: $\bar{d} = (w_1, w_2, ..., w_N)$. Accordingly, the hieroglyphic dictionary of the summary r consists of a subset of the dictionary of the document and the summary can also be put in line with the vector of weight values: $\bar{r} = (\hat{w}_1, \hat{w}_2, ..., \hat{w}_N)$. In this case, we give a natural definition:

$$\hat{w}_i = \begin{cases} w_i, & if \quad t_i \in r; \\ 0, & if \quad t_i \notin r. \end{cases}$$

It is known that the scalar product of two nonzero vectors in Euclidean space A and B is defined by a formula:

$$\bar{A} \cdot \bar{B} = \left\|\bar{A}\right\| \left\|\bar{B}\right\| \cos\theta$$

Here θ – a corner between the considered vectors. It is natural if the direction of vectors coincides, the value θ becomes equal to zero (respectively, $\cos\theta = 1$). I.e. than closer $\cos\theta$ to unit, the direction of vectors is closer to those that is easily substantially interpreted for a case of the document and its summary (the short summary). It is accepted function of proximity between vectors A and B to designate as $Sim(\bar{A}, \bar{B})$ (from word Similarity). In case of studying of a cosine measure of proximity we have:

$$Sim(\bar{A}, \bar{B}) = \cos\theta = \frac{\bar{A} \cdot \bar{B}}{\left\|\bar{A}\right\| \left\|\bar{B}\right\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

where A_i and B_i – components of vectors $\bar{A}$ and $\bar{B}$, respectively.

According to definition of a cosine measure for calculation of proximity of the document and the paper it is possible to use a formula:

$$Sim(d, r) = \frac{\sum_{i=1}^{N} w_i \hat{w}_i}{\sqrt{\sum_{i=1}^{n} w_i^2} \sqrt{\sum_{i=1}^{n} \hat{w}_i^2}},$$

Proceeding from the fact that $\sum_{i=1}^{N} w_i \hat{w}_i = \sum_{i=1}^{N} \hat{w}_i \hat{w}_i$, we receive a formula which in is used at practical calculations:

$$Sim(d,r) = \frac{\sum_{i=1}^{N} \hat{w}_i \hat{w}_i}{\sqrt{\sum_{i=1}^{n} w_i^2} \sqrt{\sum_{i=1}^{n} \hat{w}_i^2}} = \sqrt{\frac{\sum_{i=1}^{N} \hat{w}_i^2}{\sum_{i=1}^{N} w_i^2}}.$$

Another used criterion for formal identification of the degree of closeness – Jensen-Shannon divergence, which is based on the formalism of information theory and mathematical statistics, in particular, on the relative entropy of Kullback-Leibler [9], [10].

The Kulbak-Leibler entropy is generally defined as a non-negative functional, which is an asymmetric measure of the distance between two probability distributions defined on a common space of elementary events.

The divergence distribution Q relatively P is designated $D(P\|Q)$. Distribution Q often serves as distribution P approach. This measure of distance in the theory of information is also interpreted as the size of losses of information when replacing true distribution P to distribution Q. The functional value can be understood as the number of unaccounted information of distribution Q if it was used for approach the distribution P.

For discrete probability distributions P $\{p_1, p_2, ..., p_n\}$ and Q $\{q_1, q_2, ..., q_n\}$ the Kulbak-Leibler's entropy is defined as follows:

$$D(P\|Q) = \sum_{i=1}^{n} \log \frac{p_i}{q_i} p_i.$$

Kulbak-Leibler's entropy, substantially close to the concept of distance, could be called a metric in the space of probability distributions, but this would be incorrect, since it is not symmetrical $D(P\|Q) \neq D(Q\|P)$ and does not satisfy the inequality of the triangle. In the future, we will use Jensen-Shannon divergence (JSD), which is based on Kulbak- Leibler's entropy, but is a metric [18], [5], so it is also called "Jensen-Shannon's distance" [6], [7], [13].

Jensen-Shannon's divergence is defined as follows:

$$JSD(P\|Q) = \frac{1}{2}\left(D(P\|M) + D(Q\|M)\right),$$

where $M = \frac{1}{2}(P+Q)$.

In case of application of Jensen-Shannon's distance to a problem of assessment of quality of summaries s the number of the lost information in the summary in comparison with the initial document is estimated. As well as in a cosine measure, it is supposed that to the document d there corresponds the vector of the hieroglyphs weights $\bar{d} = (w_1, w_2, ..., w_N)$, and to the summary r – a vector of weight values: $\bar{r} = (\hat{w}_1, \hat{w}_2, ..., \hat{w}_N)$. The "average" vector used in Jensen-Shannon's method is presented in the following form:

$$\bar{M} = \frac{1}{2}\left(\bar{d} + \bar{r}\right).$$

Respectively,

$$JSD = \frac{1}{2}\left(D(\bar{d}\|\bar{M}) + D(\bar{r}\|\bar{M})\right) = \frac{1}{2}\left(\sum_{i=1}^{N}\log\left(\frac{w_i}{\frac{1}{2}(w_i+\hat{w}_i)}\right)w_i + \sum_{i=1}^{N}\log\left(\frac{\hat{w}_i}{\frac{1}{2}(w_i+\hat{w}_i)}\right)\hat{w}_i\right).$$

Let's consider the given sums on two areas of index values: the 1st area where hieroglyphs of the document and the summary coincide and 2nd, where do not coincide, i.e. where $\hat{w}_i = 0$:

$JSD = JSD_1 + JSD_2$.

In the first area, obviously,

$$JSD_1 = \frac{1}{2}\sum_{i=1}^{N}\log\left(\frac{w_i}{\frac{1}{2}(w_i+w_i)}\right)w_i + \frac{1}{2}\sum_{i=1}^{N}\log\left(\frac{w_i}{\frac{1}{2}(w_i+w_i)}\right)w_i = 0.$$

In the second area, respectively,

$$JSD_1 = \frac{1}{2}\sum_{i=1}^{N}\log\left(\frac{w_i}{\frac{1}{2}w_i}\right)w_i + \frac{1}{2}\sum_{i=1}^{N}\log\left(\frac{\hat{w}_i}{\frac{1}{2}(w_i)}\right)\cdot\hat{w}_i \to \frac{1}{2}\sum_{i=1}^{N}w_i.$$

Strictly speaking, the second summand in the latter formula is not correct (you can consider the limit of expressions under the sign of the

sum of when $\hat{w}_i \to 0$), but at the same time, we can make a fairly obvious conclusion that the Jensen-Shannon measure corresponds to the loss of information when summarizing and proportional to the total weight of the words (in our case – characters) included in the document, but missing in the summary.

Comparison of methods

When summarizing the new idea of determination of weight values of sentences on the basis of weights of separate hieroglyphs, but not words as it is standard was realized. Therefore the quality of summarizing is checked not only proceeding from accounting of scales of separate hieroglyphs, but also taking into account scales of the whole words included in the documents and summaries to be convinced that the offered approach is satisfactory also by criteria of traditional systems of summarizing. Naturally, this had to perform resource-intensive procedure of segmentation of words [1]. It should be noted that this procedure was performed only for quality check of algorithms of summarizing and is not a part of these algorithms.

The tests were conducted on a real array of legal information of the People's Republic of China in the amount of 10 thousand documents.

In Fig. 3-6 the results of the conducted tests are shown. In Fig. 3 and 5 are the results, when the models of documents and summariescorresponded to vectors, elements of which - weights of individual hieroglyphs from the text of the document by $TF \cdot IDF$. In Fig. 4 and 6 – results, when elements of vectors correspond to weight values of words, segmented from texts of documents and summaries. In Fig. 4 and 6 the results are given in accordance with the cosine measure of the proximity of the document and the summaries, and in Fig. 5 and 7, according to the Jensen-Shannon's distance.

On the horizontal axis on all figures the number of offers included in the summary is marked. The vertical axis shows the values of the corresponding criteria, which are averaged throughout the document array. It should be noted that in all examples, as the first sentence of the summary includes the title of the document, so the values with argument 1 for all four types of algorithms ($\sum tf \cdot idf$, Nearest, Network, Random) are the same.

The test results allow to summarize:

1. The proposed approaches lead to results, with quality is not lower then presented at the well-known conference on the analysis of texts TAC [Lois, 2008].
2. If the criterion of cosine measure of the proximity of the document

and the summary when taking into account the weight values of the individual hieroglyphs, the best results showed the method $\sum tf \cdot idf$ (which, of course, on the sum $TF \cdot IDF$ determined the weight of proposals, with the most significant included in the summary), then by the same criteria, the proposed network method was the best way to take into account separate words of natural language.

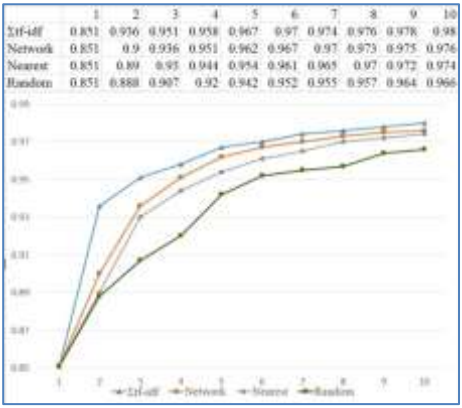


Figure 3. A cosine measure of proximity of the text and summary – accounting of the weight values of separate hieroglyphs

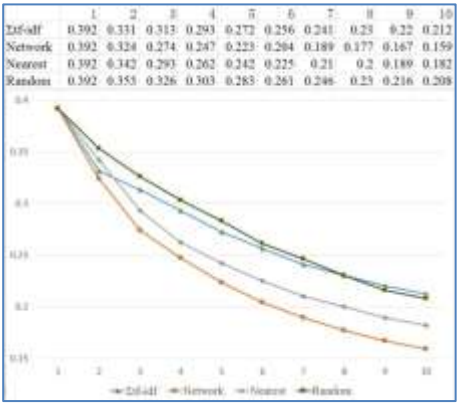


Figure 4. Jensen-Shannon's divergence of loss of information when summarizing – accounting of weight of separate hieroglyphs

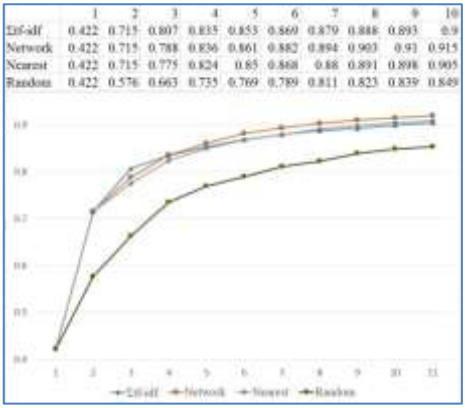


Figure 5. A cosine measure of proximity of the text and abstract – accounting of the weight values of separate words

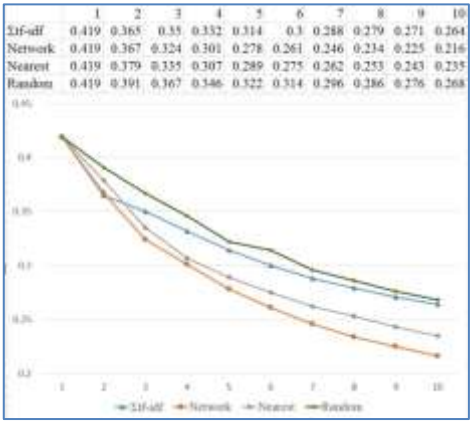


Figure 6. Jensen-Shannon's divergence of loss of information when summarizing – accounting of weight values of separate words

Conclusions

1. We introduce a new hybrid method for automatic text summarization, covering statistical and marker methods, as well as taking into account the location of sentences in the text of the document. The offered model of the paper summaryt reflects information need of customers during the work with legal information.
2. We brought the approach to determination of weights of separate hieroglyphs instead of segmented words in the text of documents.

This technique avoids the expensive procedure of words segmentation required for other semantic methods of Chinese language processing.

3. Various methods of automatic text summarization are implemented and tested. Summarizing on the basis of the offered network model of the document was the best by criteria of a cosine measure and Jensen-Shannon's distances for papers which volume exceeds 2 sentences.
4. The offered approach, with minimal modifications, can be applied to texts on any subject of scientific, technical or news information.

References

1. *Berezin, Boris A.; Lande, Dmitry V.; Pavlenko, Oleh Y.* (2017). Development, Evaluation and Usage of Word Segmentation Algorithm for National Internet Resources Monitoring Systems. CEUR Workshop Proceedings. Selected Papers of the XVII International Scientific and Practical Conference on Information Technologies and Security (ITS 2017). **2067**:16-22.
2. *Bharti, Santosh Kumar; Babu, Korra Sathya; Pradhan, Anima* (2017). "Automatic Keyword Extraction for Text Summarization in Multi-document e-Newspapers Articles". European Journal of Advances in Engineering and Technology, **4** (6): 410-427.
3. *Cherven, Ken* (2013). "Network Graph Analysis and Visualization with Gephi". Packt Publishing. ISBN: 9781783280131
4. *Chien, L.-F.* (1997). "Pat-tree-based keyword extraction for Chinese information retrieval". ACM SIGIR Forum. 31, ACM, pp. 50-58.
5. *Dagan, Ido; Lillian Lee; Fernando Pereira (1997).* "Similarity-Based Methods For Word Sense Disambiguation". Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics: 56–63. ArXiv:cmp-lg/9708010. DOI: 10.3115/979617.979625.
6. *Endres, D. M.; J. E. Schindelin (2003).* "A new metric for probability distributions". IEEE Trans. Inf. Theory. **49** (7): 1858–1860. DOI: 10.1109/TIT.2003.813506.
7. *Fuglede, Bent; Topsøe, Flemming (2004).* "Jensen-Shannon divergence and Hilbert space embedding". Proceedings of International Symposium on Information Theory, ISIT 2004, p. 31.
8. *Kleinberg, J.* (2000). "Navigation in a small world". Nature, **406** (6798): 845. DOI: 10.1038/35022643
9. *Kullback, S. (1959), Information Theory and Statistics, John Wiley & Sons.* Republished by Dover Publications in 1968; reprinted in 1978: ISBN 0-8446-5625-9.
10. *Kullback, S.; Leibler, R.A.* (1951). "On information and sufficiency". Annals of Mathematical Statistics, 22 (1): 79-86. DOI: 10.1214/aoms/1177729694.

11. *Lande, D.V.; Snarskii, A. A; . Yagunova, ; Pronoza E. V.* (2013). “The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text”. 12th Mexican International Conference on Artificial Intelligence. pp. 209-215. DOI: 10.1109/MICAI.2013.33
12. *Langville, Amy N.; Meyer, Carl D.* (2011). “Google's PageRank and beyond: the science of search engine rankings”. Princeton university press. ISBN: 9780691152660
13. *Lin, J. (1991).* "Divergence measures based on the shannon entropy". *IEEE Transactions on Information Theory, **37** (1): 145–151. DOI:* 10.1109/18.61115.
14. *Louis, Annie; Nenkova, Ani* (2008). “Automatic Summary Evaluation without Human Models”. In First Text Analysis Conference (TAC'08), Gaithersburg, MD, Etats-Unis, 17-19 November 2008.
15. *Luhn, Hans Peter* (1958). “The automatic creation of literature abstracts”. IBM Journal of research and development, 2:159–165.
16. *Ramos, J.* (2003). “Using tf-idf to determine word relevance in document queries”. Proceedings of the first instructional conference on machine learning, pp. 1-4.
17. *Salton, G.; Buckley, C.* (1988). “Term-weighting approaches in automatic text retrieval”. Information Processing & Management, **24**(5): 513—523.
18. *Schütze, Hinrich; Manning, Christopher D. (1999).* Foundations of Statistical Natural Language Processing. *Cambridge, Mass: MIT Press. p. 304.* ISBN 0-262-13360-1.
19. *Yatsko, V.A.* (2002). “Symmetric Summarization: Thematic Foundations and Methods”. Nauchno-Tekh. Inf., Ser. 2. – N. 5: 18–28.
20. *Zhang, C.* (2008). “Automatic Keyword Extraction from Documents using Conditional Random Fields”. Journal of Computationa

HARDWARE DATA ENCRYPTION COMPLEX BASED ON PROGRAMMABLE MICROCONTROLLERS

Tmienova Nataliia[1] and Sus' Bogdan B.[2]
[1]Faculty of Information Technology
Taras Shevchenko National University of Kyiv, Ukraine
[2]Institute of High Technologies
Taras Shevchenko National University of Kyiv, Ukraine
tmyenovox@gmail.com,bnsuse@gmail.com

The growing danger of computer crime puts forward a new set of urgent tasks. At the same time, the development of hardware encryption systems can be effective in the solving some of them. To study the hardware possibilities of reducing the probability of unauthorized access to information, an available complex for demonstrating data encryption based on programmable microcontrollers is proposed. The article provides a diagram of the corresponding demonstration complex with a description of its work.

The algorithms and methods of information protection between devices using combined communication channels and embedded systems are offered. The software-hardware complex, which uses a combination of engineering and software solutions, provides the possibility to conduct spectral analysis of data streams and research noise and signals that may occur in optical communication lines and radio channels.

In addition, this complex can be used for laboratory work on cryptography and data protection in communication lines.

Keywords: *cryptography, hardware encryption systems, programmable microcontrollers, STM32, security, privacy, forensics analysis. embedded systems.*

Introduction

Currently, a growing trend of data security disturbance is being observed. Information leaks often occur as a result of poor security management or as a result of using outdated or improperly implemented security procedures and technologies. Data encryption using appropriate key management schemes can significantly reduce data leakage.

However, when we use the key methodology, the following main problems arise: the generation and safe transfer of keys to interaction participants; setting of a secure information transmission channel between

the interaction participants before the transfer of keys; authentication. There are two key methodologies: symmetric (with a secret key) and asymmetric (with a public key). Each methodology uses its own procedures, its own key distribution methods, key types, as well as key encryption and decryption algorithms [1].

Although cryptography with well-known standards, modern algorithms and libraries is quite effective, the development of hardware encryption systems is still an urgent task [2, 3]. Information security software tools are potentially vulnerable, since the entire process of encoding (encrypting/decrypting) data is performed in the internal computer memory, which can be accessed by any application running on the computer.

This means that it is possible to conduct multi-level attacks on any software, including the one aimed at ensuring the security of the information. Thus, it is practically impossible to build a high-level protection by software tools only [4]. To restrict access to means performing cryptographic transformations, it is necessary to transfer them from a computer to a closed hardware subsystem.

As a result, the attacker will not be able to access directly the encoding processes (encryption/decryption). Firstly, the hardware implementation of the encryption algorithm guarantees the invariance of the algorithm itself, whereas the software algorithm can be intentionally modified. In addition, the hardware encoder eliminates any interference in the encoding process.

Another advantage is the use of a hardware random number generator, which guarantees absolute randomness of the generation of encryption keys and improves the quality of the implementation of various cryptographic algorithms. In addition, the hardware encoder allows you to load directly encryption keys into the encryption processor, bypassing the computer's RAM, while in the software encoder the keys are in memory even during the operation of the encoder. It is important that it is possible to create various systems for distinguishing and restriction of access to computer on the base of the hardware encoder. Also, the use of hardware systems makes it difficult to conceal evidence of interference in the communication channel [4].

In this paper, to investigate the hardware capabilities of reducing the probability of unauthorized access to information, an accessible data encryption complex based on programmable microcontrollers is proposed. This embedded system allows the combined use of optical and wireless communication channels.

Optic-fiber communication lines allow the information transfer over long distances with minimal distortion, which allows improving the protection technology of information transmission in optical communication lines from the harmful effects of intruders due to unauthorized connection.

The use of polarization modulation of light in the hardware complex causes the additional interest. With the help of optical information of processing devices, the decomposition of optical signals is performed by a given system of functions. The work of such devices is based on the application of electronically controlled anisotropic environment that change the polarization of the light beam. This allows to process signals.

This complex can also be used to successfully students' study of technologies, algorithms and physical methods of encryption and secure transmission of signals in communication channels, hardware capabilities of reducing the probability of unauthorized access to information upon data leakage during laboratory work.

1 Complex description

To evaluate the effectiveness of data encryption algorithms in optic-fiber systems, a number of practical solutions that include software codes and hardware implementation based on embedded systems have been developed. Optical fiber, preferably, does not have protection from third-party connections and listening. At present, a large amount of critical information is transmitted through the optical channel, and there is the risk that it may get to the intruders who has the necessary resources and equipment. Connection to optic fiber is a process in which the security of an optical channel is disturbed by the flow or leakage of light information. Therefore, a combination of standard optical channels with the radio channels, which are switched by a special algorithm, is offered. This complicates the information leak when connecting of intermediate device.

Nowadays, there is a need for new effective ways of encryption keys forming. Additionally, the access key can be transmitted using RFID cards. The possibility of using a hardware random number sensor, which produces a statically random and unpredictable signal that is converted then into a digital form, guarantees the chance of encryption keys generating and improves the quality of the various cryptographic algorithms implementation.

The proposed encryption does not eliminate the possibility of data intercepting through the optical channel, but renders stolen information inappropriate for intruders. There are high requirements to RFID systems,

since these systems are very tiny with power from the electromagnetic field.

The complex is based on the high-performance microcontroller STM32 [5, 6]. To connect the transmitters, receivers and the computer, the device interface is used. The information encryption complex has the following basic functions:

- transmission and reception of signals by separate optical channels;
- transmission and reception of signals over a common optical channel with the use of spectral multiplexing;
- transmission of signals or encryption key via radio channel with the use of digital noise filtering algorithms;
- the possibility of synchronous switching of communication channels.

The diagram of the demonstration complex is shown in Figure 1.

A modified VirtualWire open source library for STM32CubeMX [7] is used for radio modules functioning. For encoding signals, STM32 library for AES Encryption and x-cube-cryptolib libraries are implemented, which support such encryption algorithms as AES-128, AES-192, AES-256, ECB (Electronic Codebook Mode), CBC (Cipher-Block Chaining), CTR (Counter Mode) CFB (Cipher Feedback), OFB (Output Feedback), CCM (CBC-MAC), GCM (Galois Counter Mode), CMAC, KEYWRAP, XTS); hash functions supported by HMAC (SHA-1, SHA-224, SHA-256, SHA-384, SHA-512) [8, 9].

The devices transmit a message, which can be additionally encrypted by the software and hardware means. When software encryption is used, all types of information are splitted into small fixed-length packages containing special headers (so-called cells). At the next step, the spectral multiplexing in the digital channel is implemented. In this case, the cells may have different priorities.

The message is entered into the terminal program window of the STM32 module that is used as the transmitter. For the evidence in the case of demonstration mode of transmitting information through an optical channel, red, green, and blue LEDs can be used. Each color of the LED symbolizes a separate data transmission channel.

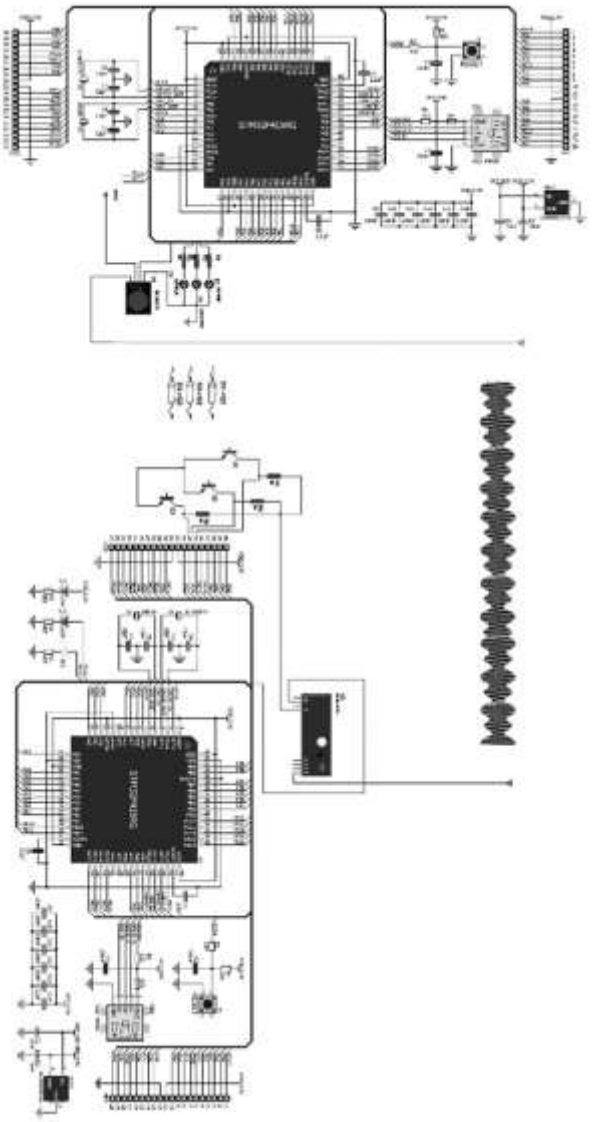


Fig.6**.** Diagram of the demonstration complex.

For transmission of data, the radio channel on available modules MX-F01 and MX-RM-5V is used. The module can be implemented to RC-control of models, security and automation projects, and wireless sensors systems construction up to 150 m [10]. The connection between the radio modules and the microcontroller is organized through the peripheral interface of the USART (Universal Synchronous/Asynchronous Receiver-Transmitter) in UART5 mode operation. The transmitting module is connected to a PC12 pin configured to data transmission via the USART, and the PD2 output configured as receiver.

The number of LEDs and radio modules can be scaled according to the number of data channels. Additionally, it is possible to use RFID labels to change configuration settings and encryption key. During transmission of encrypted message, synchronous switching of channels is implemented (switching of transmitters and receivers according to a certain encryption algorithm) (Fig. 2). The reception efficiency varies depending on the transmission speed and the delay between the packets. In this demonstration device, a combination of optical channels with spectral multiplexing of signals, combining multiple channels into one stream, use of delays, transmission of redundant information, noise overlays and a hopping code algorithm can be also used.

In the radio channel, the function of the transformation in the frequency band is also possible. With the change of carrier frequency for different signals, each of them can be transferred to another frequency band using a cyclic inversion.

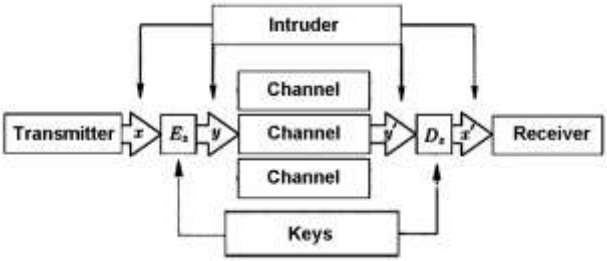


Fig.2. Switching of receiver and transmitter (x - data packets, y - encrypted data, Ez -encryption algorithms).

As the use of radio transmitters and receivers with digital frequency synthesis, the implementation of the hopping frequency tuning is possible. As an optical fiber, an OD AV Optical Fiber Cable of type Toslink can be

used. The decrypted message is displayed in the serial port terminal program window that receives data from the STM32 controller, which operates in receiver mode. This encoding is especially effective for the transmission of short data packets.

Additional encryption of information in the optical channel is possible by changing the polarization. The use of polarized light allows to realize the optical representation of alternating functions or matrices and to carry out both direct and reverse transformations in one cycle, simultaneously obtaining in the initial plane of the corresponding combination of light fields by polarization and intensity. The use of polarization cells on liquid crystals and analyzers varies according to encryption algorithm. The low speed of information transmission in this mode is limited with the use of low-cost electromechanical systems for polarizer and analyzer devices. The polarization modulation provides the optical realization of direct and reverse transformations in the communication line, as well as the separation of channels in the digital transmission of data. The software solution can be easily adapted for 32- and 64-bit microcontrollers with architecture ARM7 and ARM9.

2 Use of the complex in the “Internet of things”

The complex is designed to select the optimal multi-channel technology for data encryption in microcontroller systems with limited memory. Currently, computer systems, including intelligent systems, automated and telecommunication systems for managing things consist of a wide range of embedded devices. Especially it concerns the so-called "Internet of Things," which is a wireless network that is self-configurable between objects of different classes.

The wide implementation of embedded systems requires an integrated combination of software and hardware solutions for the protection of information both in the computing systems and in communication channels. Such devices can often store some amount of indirect information about the behavior of their owners. Information can be obtained from the analysis of various received signals, even in analog form, from the objects of monitoring and from signals of control to these objects.

For example, indirect information can be obtained from the analysis of intercepted signals from monitoring objects and sensors, usually in analog form.

The outgoing control of these objects in the form of discrete and continuous signals, as well as the transformation of analog signals in

digital and vice versa during the digital transmission of data through standard interfaces can be also intercepted.

In “Smart Home” systems, for instance, most of devices are electronically controlled. Any controlling device or computing device can potentially be used for criminal activity, so it is highly desirable to split this information into parts and apply the encryption. (For example, the stream of video data from CCTV cameras or videoregistrators, it is desirable to split into separate encrypted channels and transmit through the combined lines of communication in order to reduce the probability of interception or falsification of data). But the most important task on the “Internet of things” is to ensure the security of the main and additional communication channels. Implementation of optical and wireless interfaces for connecting the main embedded devices to "Smart Home" systems and the use of combined communication channels will greatly increase the security of data transmission. The intruder must have physical access to all memory chips and use reverse engineering technology to obtain low-level information from the flash memory to get some data only about the encryption algorithm. Due to the implementation of the multichannel data transmission circuit, the CPU load is also reduced. The complex provides an opportunity to carry out modeling in the engineering design of prototypes of a secure system for the Internet of things and to investigate the technical characteristics of the expected equipment. It is possible to conduct a comparative testing of the security of data of various devices, depending on the used channel, the algorithm and the key of encryption and communications. Since the encryption key can be transmitted over a channel that is randomly selected and can be split into parts by the special algorithm, this significantly complicates the access of potential intruder and analysis of the encrypted information. Another advantage of the proposed system is the possibility of scaling the number of communication channels without significant increase of the costs and the ability to support various encryption protocols without major reconfiguration of the system.

3 Conclusions

The growing danger of computer crime poses a set of new topical issues. At the same time, the development of hardware encryption complexes can be effective on the way to overcome some of them.

The discussed demonstration complex allows us to assess the efficiency of encrypting data streams in channels, compare packets received from the transmitter with the number of sent packets, analyze the

spectrum of the encrypted signal and the radio channel noise by the means of computer graphics techniques.

This complex can be successfully implemented in the following areas: banking, military and medical industries, telephony. The main advantages include the reliability of transmission, the simplicity of implementation, the flexibility of the functionality and the possibilities of application and modernization.

We also note that the demonstration complex can be used for conducting laboratory classes for the development of protocols for the transmission of information and digital signal processing filters analysis.

The complex makes it possible to conduct physical experiments to monitor signals in an optical fiber and a radio channels for the choice of an optimal algorithm for encryption stability and simulation for coherent and incoherent optical processors. The proposed approaches can be used to modify data transmission equipment and evaluate the reliability of encryption. It can be also used for the reverse engineering demonstrations.

References

1. Introduction to Cryptography - authorized translation of an article by J. Chandler "Cryptography 101" (in Russian) [Electronic Resource]. Mode of access: http://citforum.ck.ua/security/cryptography/crypto_1.shtml
2. Security of information systems (in Russian) [Electronic Resource]. Mode of access: http://intuit.valrkl.ru/course-1312/index.html.
3. Security with STM32 & Secure Elements [Electronic Resource]. Mode of access: http://www.emcu.it/SILICA-STDay2016/X/Presentazioni/2_STM32&SecureElements.pdf
4. Stallings W. Cryptography and network security: principles and practice. – New York: Prentice Hall, – 2006. – 680 p.
5. Cortex-M Series from ARM.com [Electronic Resource]. Mode of access: http://www.arm.com/products/processors/cortex-m.
6. Cortex-M4 Technical Reference Manual. [Electronic Resource]. Mode of access: http://infocenter.arm.com/help/topic/com.arm.doc.ddi0439b/DDI0439B_cortex_m4_r0p0_trm.pdf
7. VirtualWire library for Arduino and other boards [Electronic Resource]. Mode of access: https://www.airspayce.com/mikem/arduino/VirtualWire/

8. STM32 cryptographic library (UM0586)) [Electronic Resource]. Mode of access: https://www.st.com/en/embedded-software/STM32-cryp-lib.html
9. STM32 crypto library (UM1924) [Electronic Resource]. Mode of access: https://www.st.com/content/ccc/resource/technical/document/user_manual/group0/f9/6e/f2/a2/b4/ec/49/c0/DM00215061/files/DM00215061.pdf/jcr:content/translations/en.DM00215061.pdf
10. Complete Guide for RF 433MHz Transmitter/Receiver[Electronic Resource] Mode of access: https://randomnerdtutorials.com/rf-433mhz-transmitter-receiver-module-with-arduino/

АНАЛІЗ ПРОЦЕДУРИ ОЦІНЮВАННЯ СТАНУ КІБЕРВРАЗЛИВОСТІ СИСТЕМ ЕЛЕКТРОПОСТАЧАННЯ

Ігор Яковів[1], Віталій Циганок[2]

[1] Інститут спеціального зв'язку та захисту інформації НТУУ "КПІ імені Ігоря Сікорського" , Київ, Україна, iyakov52@gmail.com

[2] Інститут проблем реєстрації інформації НАН України , Київ, Україна, tsyganok@ipri.kiev.ua

Зростають інтенсивність застосування, різноманітність та результативність складних кібератак на національні системи електропостачання. Цьому процесу сприяє постійне розширення знань зловмисників про вразливості комп'ютерних систем и вдосконалення технологій втручання. Існуючі концепції щодо оцінювання стану безпеки систем інформаційних технологій не орієнтовані на оперативне врахування змін факторів, що обумовлюють результативність захисту. Кібернетичний сегмент систем електропостачання має свої особливості, врахування яких може підвищити ефективність оцінювання.

Пропонується підхід до вирішення задачі розробки процедури оперативного оцінювання стану кібербезпеки систем електропостачання, який основується на знаннях про рівень актуальних загроз і особливості процесів управління цими системами. У роботі представлені результати аналізу компонентів такої системи оцінювання, визначені основні принципи її функціонування.

Ключові слова: *Кібербезпека, кібервразливості, експертне оцінювання, підтримка прийняття рішень, системи електропостачання.*

Вступ

Серія кібератак на системи управління інфраструктури електропостачання України була зафіксована у 2015-2016 роках. Відключення 23 грудня 2015 року 225000 клієнтів ПАТ "Прикарпаттяобленерго" стало першою у світі зареєстрованою успішною кібератакою на енергетичну систему з виведенням її із ладу. Системи електропостачання є важливою компонентою національної критичної інфраструктури. Після цих подій для багатьох фахівців у сфері національної безпеки різних країн важливим стає питання про можливість повторення аналогічного сценарію у своїй країні. В комплексі заходів захисту від кібератак інфраструктури електропостачання ключове місце займає процедура

оцінювання актуального стану кібербезпеки (далі – Оцінювання). За результатами цієї процедури фахівцю зазвичай потрібно отримати інформацію (оцінку), яка дозволяє йому відповісти на питання:

- Чи відповідає (не відповідає) стан захисту рівню актуальних кіберзагроз ?
- В якій мірі стан кібербезпеки не відповідає необхідному стану?
- Які заходи потрібно виконати для досягнення необхідного стану?

У свою чергу, процедуру Оцінювання можливо представити наступними пов'язаними процедурами:

- отримання інформації про поточний стан об'єкту оцінювання (ОО);
- нормування отриманої інформації відповідно до шкали оцінювання (ШО);
- формування оцінки на основі порівняння із критерієм (-ями) оцінювання.

В умовах зростаючої інтенсивності застосування нових складних кібератак на критичну інфраструктуру, що базуються на постійному розширенні знань зловмисників про вразливості комп'ютерних систем, актуальною стає задача розробки оперативних методів оцінювання стану кібербезпеки систем електропостачання.

Для подальшого наповнення конкретним змістом Оцінювання, необхідно уточнити наступні аспекти:

- що представляє собою об'єкт оцінювання (ОО);
- які сучасні принципи забезпечення кібербезпеки застосовуються;
- яким чином можливе використовування існуючих підходів оцінювання безпеки систем інформаційних технологій (IT-systems) до оцінювання стану кібербезпеки систем електропостачання.

Мета досліджень: визначення, аналіз і формалізація факторів, що впливають на систему оперативного оцінювання стану кібербезпеки систем електропостачання.

1 Аналіз об'єкту оцінювання

Сучасні електроенергетичні системи (англ. - *power grids*) мають дві основні складові:

- енергетична компонента (генерація електроенергії, розподіл, доставка, забезпечення клієнтів);

– кібернетична (комп'ютерна) компонента (моніторинг енергетичних процесів і керування їх станом).

До складу енергетичної компоненти відносяться:
– електростанції (генеруючі об'єкти),
– електричні підстанції (перетворення електроенергії за допомогою трансформаторів та інших пристроїв; розподіл електропотоків за допомогою комутуючих пристроїв; контроль за станом електропотоків за допомогою різноманітних сенсорів і актуаторів);
– мережі електропостачання, що поєднують електростанції із підстанціями і клієнтами (споживачами);
– абонентські пристрої клієнтів.

Кібернетичну компоненту прийнято називати *індустріальною системою управління* (Industrial Control System, ICS). В свою чергу в залежності від масштабів і завдань управління ICS поділяються на:
– системи диспетчерського управління та збору даних (Supervisory Control and Data Acquisition (SCADA) systems);
– розподілені системи управління (Distributed Control Systems, DCS);
– програмовані логічні контролери (Programmable Logic Controllers, PLC).

PLC можуть входити до складу DCS. У свою чергу, DCS можуть входити до складу SCADA систем.

Для управління бізнес-процесами оператора електропостачання використовуються окремі системи інформаційних технологій (далі, IT-system of management, ITSoM). З міркувань кібербезпеки такі системи повинні бути відокремлені від ICS [1], але на практиці це часто порушується. Відомі кібератаки на системи електропостачання України було здійснено через ITSoM. У подальшому при визначені сутності оцінювання стану кіберврразливості це необхідно враховувати. Наприклад, при уточнені складу об'єкту оцінювання потрібно поділяти ситуації входження/не входження ITSoM до кібернетичної компоненти power grid. Доцільно на першому етапі проводити оцінювання для ситуації "кібернетична компонента - тільки ICS ", а на другому - "кібернетична компонента - ICS і ITSoM".

2 Аналіз актуальних принципів забезпечення кібербезпеки

Аналіз наявних практик захисту корпоративних інформаційно-телекомунікаційних систем (ІТС) дозволяє виділити дві актуальні стратегії протидії кібератакам: реактивний захист і проактивний (превентивний) захист. Загальною основою для цих стратегій є процеси:

1. спостереження в реальному часі (in real time) за подіями в позначеному сегменті кіберпростору;
2. формування за допомогою сенсорів, збір і нормування інформації про події безпеки в єдиному центрі оперативної обробки;
3. аналіз подій і прийняття рішення про наявність кібератаки;
4. прийняття рішення про протидію атаці і реалізація цього рішення за допомогою актуаторів безпеки (виконавчих пристроїв безпеки).

Для реактивної стратегії прийняття рішення про виявлення атаки відбувається після її закінчення. Заходи протидії можуть запобігти лише такій же самій наступній атаці. Для проактивного (превентивного) захисту виявлення атаки має відбутися ще до її завершення. У такому випадку залишається час на реалізацію заходів переривання цієї атаки.

Ключовою частиною сучасних систем кіберзахисту корпоративних ІТС є центри операцій кібербезпеки (the CyberSecurity Operations Center, CSOC або SOC). Такі центри за допомогою операторів і/або засобів управління інформацією і подіями безпеки (Security of Information and Event Management, SIEM) з різним ступенем автоматизації реалізують перераховані вище процеси 1-4. Вітчизняні нормативно-правові документи в сфері захисту інформаційно-телекомунікаційних систем в явному вигляді не визначають і не регламентують процеси захисту в режимі реального часу. Як правило, реалізація в рамках комплексної системи захисту інформації (КСЗІ) заходів захисту від несанкціонованих дій (НСД) здійснюється на основі стратегії реактивного захисту. Інформація про події безпеки формується комплексом засобів захисту (КЗЗ) від несанкціонованих дій на основі критеріїв (ознак), які закладені ще на етапі проектування.

Реакція (відповідь) на інциденти безпеки, як правило, формується після завершення інциденту.

На рисунку 1. представлена структура і основні інформаційні процеси системи оперативного кіберзахисту.

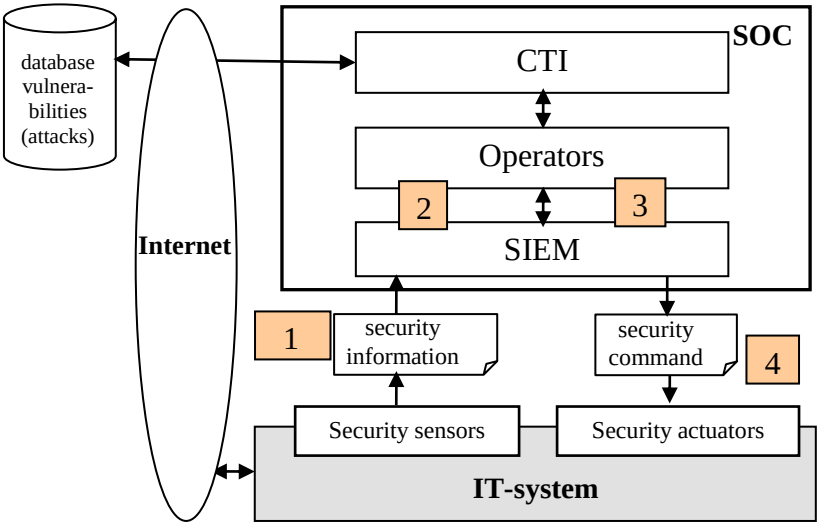


Рис. 1. SIEM – управління інформацією і подіями безпеки; CTI – розвідка кіберзагроз. Процеси: 1 – моніторинг подій безпеки; 2 – збір і аналіз; 3 – рішення про атаку: 4 – рішення про відповідь і його реалізація

Послідовне виконання процесів кіберзахисту 1-4 становить цикл, який постійно повторюється. На початку циклу сенсори безпеки (Security sensors) на основі індикаторів безпеки (індикатори компрометації, Indicators Of Compromise, IOCs) відслідковують події безпеки (компрометації). В разі визначення події інформація безпеки (security information) направляється до SOC, де оператори за допомогою програмних засобів SIEM аналізують її на відповідність політиці безпеки. У випадку виявлення кібератаки або кіберінциденту приймається рішення про підготовку відповіді. В кінці циклу актуатори безпеки (Security actuators) на основі отриманих команд безпеки (security commands) виконують це рішення. Основою моніторингу безпеки за допомогою сенсорів є IOCs, що визначаються розвідкою кіберзагроз (Cyber Threat

Intelligence, CTI). Джерелами IOCs можуть бути зовнішні бази даних загроз/атак (DataBase Vulnerabilities/Attacks, DBV/A), або спеціальні методи зовнішньої/внутрішньої CTI організації.

Широкомасштабне застосування проти національної критичної інфраструктури складних кібератак типу APT (Advanced Persistent Threat, вдосконалена/розвинена стала загроза) стало потужним стимулом для розвитку методів проактивного захисту на основі SOC.

Характерним для APTs є:

- атака представляє складний набір взаємопов'язаних за часом і простором дій зловмисника. Окремо ці дії можуть не викликати підозр;
- цільова акція атаки в кіберсегменті об'єкта готується тривалий час (від декількох місяців до року і більше);
- сукупність дій зловмисника – це ланцюжок тактик, виконання яких дозволяє досягти мети атаки. Незважаючи на різноманітність засобів, що використовуються в APTs, набір більшості тактик і їх сутність залишаються постійними.

Усі ці фактори сприяють розвитку конструктивних методів проактивного кіберзахисту від APTs.

3 Модель процедури оцінювання стану кібербезпеки системи електропостачання

Проведений вище аналіз об'єкту оцінювання дозволяє провести подальшу формалізацію. Для цього представимо процедуру оцінювання як систему у вигляді інформаційно-функціональної структури (рис.2).

Така структура є першім рівнем формалізації процедури оцінювання. Вона дозволяє конкретизувати задачу оцінювання шляхом декомпозиції на різні ситуації за певними ознаками.

Наприклад, за ознакою "склад кібернетичної компоненти" можливо отримати наступний перелік ситуацій (далі - шкала ситуацій складу кібернетичної компоненти, шкала ССКК, табл.1).

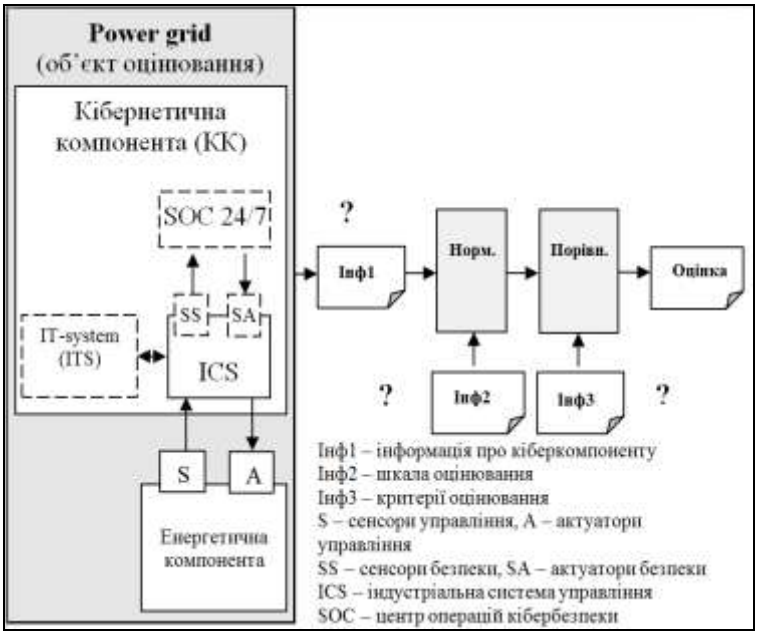


Рис. 2. Структура оцінювання стану кібербезпеки системи електропостачання

Таблиця 4.

Шкала ССКК				
	КК-1	КК-2	КК-3	КК-4
ICS	+	+	+	+
ITS	-	+	-	+
SOC 24/7	-	-	+	+

Для ситуації КК-2 в інформації про поточний стан КК (Інф1) треба відображати більше параметрів ніж для КК-1 (NКК-2 > NКК-1, де N – кількість параметрів). Для КК-3 – NКК-3 > NКК-2. Оцінювання для ситуації КК-4 буде найбільш складним за шкалою ССКК. В аспекті цієї шкали також постає питання актуальності інформації Інф1. Стрімкий постійний розвиток технологій втручання в кібернетичну компоненту ставить питання про збільшення частоти оцінювання, яку можливо представити у вигляді періоду оцінювання ТОЦ. Ситуації ТОЦ = 2 рази/рік, ТОЦ = 2 рази/місяць і ТОЦ = 2

рази/годину значно відрізняються, що впливає і на методи формування оцінки. В перших двох випадках є час для застосування методів експертної оцінки для формування Інф1 (тобто формування інформації на основі знань про актуальну ситуацію, що сформовані у компетентних спеціалістів – експертів у галузі кіберзахисту). В останньому випадку час на використання експертних підходів відсутній.

На подальшому рівні формалізації процедури оцінювання стає питання визначення сенсу (семантики) інформації, що застосовується (Інф1, Інф2, Інф3 -?). Якщо розглядати інформацію як властивість (-вості) об'єкту, що відображені у іншому об'єкті (атрибутивно-трансфертний підхід до сутності інформації, [2]), то семантикою інформації Інф2 (шкала оцінювання) буде набір параметрів безпеки, що характерні для даної ICS і були визначені заздалегідь на основі політики кібербезпеки. Семантикою інформації Інф3 (критерії оцінювання) буде набір значень параметрів безпеки, що заздалегідь визначені політикою безпеки для даної ICS. За логікою структури оцінювання, що розглядається, семантикою інформації Інф1 (поточна інформація про КК) для ситуації КК1 будуть поточні значення параметрів, що відслідковуються відповідно до шкали оцінювання (Інф2). Після нормування ці поточні значення порівнюються із критеріями кібербезпеки для даної ICS. За результатами цього формується оцінка стану кібербезпеки.

Додатково потрібно відзначити ще і про іншу необхідну інформацію, що не відображена у структурі, але її потрібно враховувати у процедурі оцінювання при формуванні. Це перелік можливих параметрів безпеки, за допомогою яких можливо формувати шкали оцінювання для ICS різної конфігурації. Назвемо таку інформацію "абеткою параметрів безпеки ICS", Інф4.

Адекватність запропонованої моделі (структури) оцінювання можливо визначити при порівнянні із існуючою системою оцінювання стану безпеки інформаційно-телекомунікаційних систем (назва, що застосовується в національних нормативних документах і аналогічна за суттю назві "IT-system"). Процедура оцінювання стану захищеності інформаційно-телекомунікаційних систем (далі – ІТС) від несанкціонованих дій регулюється документами НД ТЗІ 2.5-004-99 і НД ТЗІ 2.5-005-99 [3, 4]. Сутність існуючої процедури

оцінювання ІТС можливо коротко представити наступними чином [5]:

a) існує структурований загальний набір послуг захисту (НПЗ). Кожна послуга відповідає відомій загрозі. Послуги реалізуються відповідними засобами захисту. Всі послуги захисту структуровані на 4 групи за типами основних загроз: 1) порушення конфіденційності інформації (К); 2) порушення цілісності (Ц); 3) порушення доступності (Д); 4) порушення спостережливості (С);

b) на основі загального набору послуг НПЗ та обраної політики безпеки формується функціональний набір послуг для конкретної ІТС (далі - функціональний профіль захищеності, ФПЗ) ІТС, що відповідає основним завданням захисту;

c) відповідно до ФПЗ розробник системи захисту визначає і встановлює в ІТС сертифіковані засоби захисту;

d) оцінка стану безпеки формується в рамках державної експертизи на основі перевірки відповідності розробленої системи захисту раніше обраному ФПЗ.

В таблиці 2 представлені результати порівняння запропонованої моделі оцінювання та існуючої процедури оцінювання безпеки ІТС.

Таблиця 2.

	Процедура оцінювання безпеки ІТС	**Модель оцінювання кібербезпеки системи електропостачання**
1	Структурований загальний набір послуг захисту (НПЗ)	Абетка параметрів кібербезпеки ICS, (Інф4)
2	Функціональний профіль захищеності, ФПЗ	Шкала оцінювання (Інф2)
3	Кожна послуга із НПЗ повинна бути реалізована в системі захисту ІТС	Критерії оцінювання (Інф3)
4	Наявність послуги захисту в системі захисту (інформація визначається в рамках державної експертизи)	Поточна інформація про стан ICS (Інф1).

Кожному виду інформації, що визначена в рамках запропонованої моделі (інформаційно-функціональної структури) оцінювання стану кібербезпеки системи електропостачання

відповідає інформаційний об’єкт із існуючої процедури оцінювання стану безпеки ІТС. Відмінність процедур повинна бути встановлена в рамках додаткових досліджень із врахуванням:

- особливостей структур енергетичної (ЕК) і кібернетичної компонент (КК);
- особливості взаємодії ЕК і КК;
- вимог до оперативності процедури оцінювання для ICS.

Особливість взаємодії ЕК і КК можливо розглядати на основі моделі інформаційних процесів кібернетичної системи і критеріїв їх безпеки [6]. При цьому необхідно також враховувати підходи до формалізованого представлення кіберпростору і кібербезпеки [7].

4 Застосування технології групової декомпозиції та експертного оцінювання

У випадках, коли періодичність зміни інформації допускає можливість попереднього аналізу (не в умовах реального часу), доцільним є застосування технології, що дозволяє залучати знання експертів та інженерів по знаннях для побудови моделі предметної області. Така модель може являти собою побудовану шляхом декомпозиції ієрархію критеріїв [8], на основі якої можливо проводити оцінювання стану кібервразливості критичних інфраструктур.

4.1 Сутність технології

Технологію розроблено для застосування у слабко структурованих областях, коли інформації для обґрунтованого прийняття рішень є недостатньо і знання в області застосування суттєво обмежені, неформалізовані і здебільшого розподілені серед вузькоспеціалізованих спеціалістів-експертів. Наразі технологія реалізована у вигляді веб-орієнтованої розподіленої комп’ютерної системи [9], що дозволяє інженерам по знаннях (організаторам експертиз) та експертам не збиратись для спільної роботи у певному місці, а працювати віддалено, надаючи знання, необхідні для побудови адекватної моделі.

Технологія передбачає наступні етапи:

- Формулювання організатором експертизи головного критерію для оцінювання;
- Формування групи компетентних спеціалістів для проведення експертизи;

– Проведення колективом експертів декомпозиції головного критерію. Цей етап поділяється на наступні підетапи :
  - Формулювання кожним окремим членом групи множини факторів, що на його/її думку значно (не менше ніж на 10% від усіх впливів) впливають на формування оцінки за критерієм;
  - Коли кожен з експертів, що приймають участь у експертизі сформував множину складових критерію (найбільш суттєвих з його/її точки зору), тоді проводиться групування усіх наданих експертами формулювань у групи за однаковим змістом;
  - Проводиться груповий вибір кращого формулювання у кожній групі однакових за змістом формулювань шляхом голосування. Голосування передбачає можливість віддати голос за одне із формулювань, або відмітити своє бажання не включати жодне з формулювань групи (наприклад, через не достатньо значний вплив цього фактора на критерій, що декомпозується).

У результаті цього етапу групою експертів в умовах консенсусу формується множина найвагоміших факторів, що впливають на певний критерій. На початку – це головний сформульований критерій, а у ході подальшої декомпозиції – це може бути будь-який критерій, який доцільно розділити на складові. Таким чином формуються взаємопов'язані компоненти, які формують узагальнений критерій для оцінювання.

Декомпозиція головного критерію проводиться у результаті повторюваного ітераційного процесу, який контролюється організатором експертизи. Ця особа, як інженер по знаннях, приймає рішення про необхідність подальшої декомпозиції критеріїв, що складають множину компонентів системи оцінювання. Для кожної такої декомпозиції може формуватись окрема група експертів, найбільш компетентних у поточному питанні, що розглядається [10]. Рішення про припинення подальшого процесу декомпозиції того, чи іншого критерію приймається інженером по знаннях у тому випадку, якщо за цим критерієм можливо отримати легко-вимірюваний, бажано, кількісний показник, що характеризує систему.

– Проведення групового експертного оцінювання взаємовпливів критеріїв у системі оцінювання. У результаті цього етапу з застосуванням методів групового експертного оцінювання складами груп експертів, що проводили саме визначену декомпозицію, визначаються відносні коефіцієнти

безпосереднього впливу критеріїв на критерій вищого рівня в цій ієрархіє подібній структурі.

Після виконання цього останнього етапу модель системи оцінювання може вважатись побудованою, і тепер на основі цієї моделі, що являє собою структуру взаємопов'язаних різно-значимих факторів можливо визначити відносне значення оцінки за головним критерієм.

4.2 Можливий приклад оцінювання стану кібервразливості систем електропостачання

На гіпотетичному прикладі розглянемо застосування технології та роботу системи.

Нехай інженер по знаннях – зареєстрований в системі «Консенсус-2» користувач з наданими повноваженнями організатора експертизи сформулював головний критерій оцінювання стану кібервразливості систем електропостачання, наприклад, «Рівень захищеності кібернетичної компоненти систем електропостачання». Маючи відповідні повноваження у системі, організатор експертизи надає повноваження для роботи у якості експертів групі користувачів.

Після відповідної автентифікації кожен із експертів відповідно із своїм рівнем обізнаності сформулював перелік найбільш вагомих факторів, що впливають на рівень захищеності кібернетичної компоненти систем електропостачання. Після взаємного узгодження і узагальнення знань, наданих експертами виявились найбільш вдалими наступні формулювання критеріїв-компонентів системи оцінювання: «збереження цілісності інформації», «збереження конфіденційності інформації», «забезпечення доступності інформації». На даному етапі, інженер по знаннях мав можливість і без залучення експертних думок визначити ці компоненти системи, оскільки така декомпозиція передбачена в керівних документах щодо організації кіберзахисту [1].

Далі, кожен із критеріїв, що увійшли до згаданого вище переліку, декомпозується окремо призначеною групою експертів. У результаті отримано декомпозицію, яка у графічному інтерфейсі організатора експертизи може виглядати так, як показано на Рис.3.

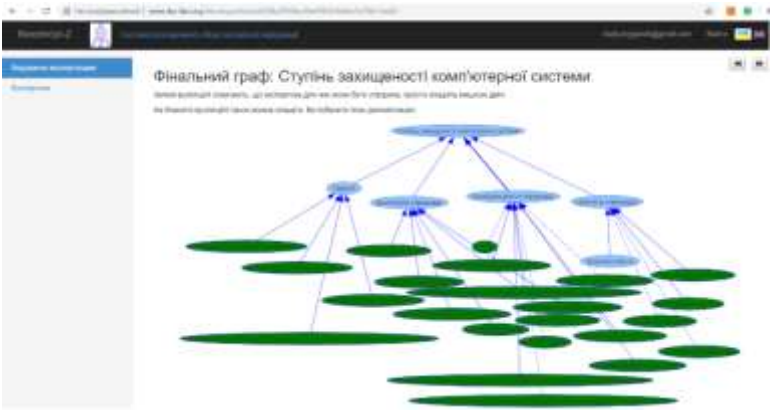


Рис. 3. Наглядне представлення створеної моделі системи оцінювання у рамках системи «Консенсус-2»

Після зважування дуг графа значеннями відносних коефіцієнтів впливу, які визначаються у результаті проведених групових експертних оцінювань, маємо ієрархію критеріїв – модель для оцінювання кібервразливостей систем електропостачання.

На нижніх рівнях ієрархії критеріїв маємо об'єктивні показники системи, які групи спеціалістів виділили, як визначальні для даного оцінювання. Маючи такі показники для наявних систем, у результаті розрахунків за методиками [11] отримаємо відносні індекси стану кібервразливостей конкретних систем електропостачання. Для випадку, коли маємо для оцінювання одну єдину систему електропостачання, то оцінювання проводиться у порівнянні зі створеним еталоном – гіпотетичної ідеально захищеної системи.

На основі таких ієрархічно-мережевих моделей можливо не тільки виконувати порівняльне оцінювання стану кібервразливостей систем електропостачання, але й виконувати оптимальний розподіл ресурсів для виконання заходів щодо поліпшення кіберзахищеності систем, будувати довготермінові плани організації кіберзахисту критичних інфраструктур [12, 13].

Висновки

У даній роботі запропоновано підхід до розробки процедури оцінювання стану кібербезпеки систем електропостачання, який базується на знаннях про рівень актуальних загроз і особливості

процесів управління цими системами. Представлено можливість застосування технології групової декомпозиції, реалізованої у рамках розподіленої системи підтримки прийняття рішень.

Література

1. Stouffer Keith, Falco Joe, Scarfone Karen Guide to Industrial Control Systems (ICS) Security. Recommendations of the National Institute of Standards and Technology. *NIST Special Publication 800-82* Rev. 2 (as of August 12, 2015) accessed on http://dx.doi.org/10.6028/NIST.SP.800-82r2 .
2. Ihor Yakoviv, "The communication channel from the position of attributive-transfer nature of the information", *Information technology and security*, vol. 1, iss. 2, pp. 84-96, 2012. Kyiv, Ukraine: Institute of special communications and information security NTUU "Igor Sikorsky KPI". [Online]. Available: http://its.iszzi.kpi.ua/issue/view/2838 .
3. НД ТЗІ 2.5-004-99 Критерії оцінки захищеності інформації в комп'ютерних системах від несанкціонованого доступу. Затверджено наказом ДСТСЗІ СБ України від 28.04.1999 № 22.
4. НД ТЗІ 2.5-005-99 Класифікація автоматизованих систем і стандартні функціональні профілі захищеності оброблюваної інформації від несанкціонованого доступу. Затверджено наказом ДСТСЗІ СБ України від 28.04.1999 № 22.
5. Яковів І.Б. Основи побудови комплексної системи захисту інформації для інформаційно-телекомунікаційної системи. Навчальний посібник. – Київ: Вид-во ІСЗЗІ НТУУ "КПІ імені Ігоря Сікорського", 2016.-88с.
6. Igor Yakoviv. "The base model of informational processes of management and safety criteria for cybernetic systems", *Information technology and security*, vol. 3, iss.**1**(4), pp.68-73, 2015. Kyiv, Ukraine: Institute of special communications and information security NTUU "Igor Sikorsky KPI". Available: http://its.iszzi.kpi.ua/article/view/57735/53977 .
7. Ihor Yakoviv. "Infocommunication system, conceptual model of cyberspace and cybersecurity". *Information technology and security*, vol. 7, iss.1(4), pp.68-73, 2017. Kyiv, Ukraine: Institute of special communications and information security NTUU "Igor Sikorsky KPI". Available: http://its.iszzi.kpi.ua/article/view/57735/53977 .
8. Saaty T.L. The Analytic Hierarchy Process: planning, priority setting, resource allocation / N.Y.: McGraw Hill. – 1980. – 287 p.
9. Свідоцтво про реєстрацію авторського права на твір №75023. Комп'ютерна програма „Система розподіленого збору та

обробки експертної інформації для систем підтримки прийняття рішень – «Консенсус-2»" / Циганок В.В., Роїк П.Д., Андрійчук О.В., Каденко С.В. // від 27/11/2017.
10. Тоценко В.Г. Методы и системы поддержки принятия решений. Алгоритмический аспект / ИПРИ НАНУ. – К.: Наукова думка, 2002. – 382с.
11. Циганок В.В. Удосконалення методу цільового динамічного оцінювання альтернатив та особливості його застосування. *Реєстрація, зберігання і обробка даних*. 2013. т.15, №**1**.– С.90-99.
12. Tsyganok V, Kadenko S, Andriychuk O, Roik P. Usage of multicriteria decision-making support arsenal for strategic planning in environmental protection sphere. *J Multi-Crit Decis Anal.* 2017; **24**:227–238.
13. Tsyganok V.V., Kadenko S.V. & Andriichuk O.V. Using different pair-wise comparison scales for developing industrial strategies. *International Journal of Management and Decision Making*. – 2015. – vol. 14, issue **3**. – P. 224-250.

DECISION SUPPORT SYSTEMS' SECURITY MODEL BASED ON DECENTRALIZED DATA PLATFORMS

Mykyta Savchenko, Vitaliy Tsyganok, Oleh Andriichuk
[1]Institute for Information Recording of National Academy of Sciences of Ukraine,Kiev, Ukraine
zitros.lab@gmail.com, tsyganok@ipri.kiev.ua, andriichuk@ipri.kiev.ua

Modern information systems, especially high-risk systems which work with important data lack proper security models. Decision support systems and recommendations they produce are extremely dependent on the data they produce recommendations on, hence they require the most secure data platform and transparent history of data input and changes. Many of such systems are centralized and are stored in a single place, like a data center or even a single machine, where important data can be lost easily. Most importantly, it can be tampered without many complications, as for the typical setup there are people who always have access to the system and its data, including people who know for sure how the system is made and who can act unnoticed, being bad actors. Through all of this can be solved by introducing proper monitoring mechanics and implementing best security practices like using secure protocols and encryption, this article describes methods which completely exclude data tampering possibility and add more important properties like data immutability, decentralization and fault tolerance on a platform level.

Keywords: *security, decentralized platforms, distributed systems, decision support system, knowledge base, data immutability.*

1 The Problem of Data Security of Decision Support Systems

Decision support systems (DSS) are systems that take facts and produce recommendations for decision-makers based on numerous factors. This recommendations is typically used for informational purposes only, however, the system's output might be considered as the primary conclusion for many complex operations that people usually cannot handle on their own (examples: social groups behavior research, a complexnetwork of facts and relationships, others) [1, 2]. The ongoing increasing of subject domains' models complexity requires an adequate and detailed representation of sets of factors and their interactions in the DSS knowledge base. Significant level of detail of knowledge base leads to redundancy, ambiguity, contradictions' presence in the base and, thus, to the deterioration of the adequacy of subject domains' models.

A structure of DSS knowledge base is shown in Fig. 1. The main elements of knowledge base are objects and connections between them. Knowledge base objects can be one of two types: target or project. Each

knowledge base object has a name in short wording form. Semantic meaning of the object is specified by a keyword tuple with corresponding weights. This object can be quantitative or qualitative, threshold or quasi-linear. Projects have specific parameters: runtime and resources required. A connection between knowledge base objects can be positive or negative, it can have a time delay, compatibility groups. It is also characterized by a relative impact factor.

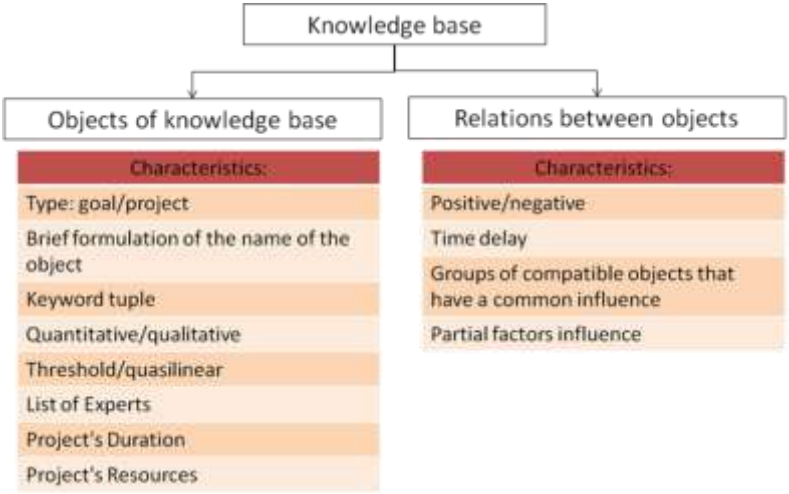


Fig. 1. Structure of DSS Knowledge Base

Because the recommendations of these systems depends on a data so much, a proper security model standard must be applied in order to prevent unauthorized or unexpected changes in the data, which may change the decision made by decision support system. There are several concepts related to security that should be reviewed.

1.1 The Need to Trust Entire System and Rely on People

In order to trust the recommendations of decision support system, one should trust the developers of this system, algorithms behind the system, people that control and maintain the system and those who interact with it and input information [2]. Furthermore, produced recommendations can be incomprehensible or undesirable, which raise more questions rather than answers regarding system operation [3]. As the result, there are a

plenty of cases in which each party have to trust the system in order to accept produced decision.

Having the decision support system that is fully under the control of one party, or even multiple parties enable them to make decisions regarding the data they own:

1. Which data to input and how to organize an input process.
2. Which data to filter from the input and which to keep.
3. How to organize a decision support process.
4. How to tune the system for a specific case.
5. How to properly present and explain the recommendations.

All of these processes rely on people who work with the system, not to mention those who input data and the system itself. In such a scheme there are too many points of failure, the main of which are people themselves. In case of undesirable results, a group of people can decide to slightly change the input data achieving the result they desire on their own. Having a system in which much of the trust relies on people is not desirable, as people never can be a source of genuine and logical decisions by their nature [1].

Instead, it is much desirable to have a system which does not require the trust to its operators and is trusted on a global level. This eliminates any human errors, as well as intentional attempts to change something in the system, leaving this change unnoticed.

1.2 Data Tampering Risks

Decision support systems are extremely sensitive to any data they store and, most importantly, take as an input, as even a slight modification in this data could drastically change the result. People who input data to the system can do errors, both accidentally or intentionally, and there is always a chance for them to play a bad role and intentionally harm the output of the system by giving an invalid or ambiguous input. Different methods can be used to reduce the input error, however, they cannot eliminate it [3].

Moreover, the input data can always be reviewed and validated by other people. Hence the only thing bad actors can perform in order to change resulting decisions is to tamper with data and hide any evidence of it.

Traditional system security models feature many ways to protect data authenticity, including encryption, data mirroring, backups, monitoring,

and logging, but even if the system is set up properly and is well-maintained, there is always a risk that something can go wrong or is missing, and more importantly there is no way to ensure that the data wasn't changed from the moment of its entry [4].

Decentralized data platforms solve this problem by introducing immutable public or permissioned ledger maintained by multiple parties that participate in a process of validating this ledger. Hence, a single entry must be trusted by all or the majority of network participants in order to be recorded on a ledger. Data is also recorded in a cryptographically secure way, excluding the possibility of previous records modification [5].

1.3 Reliability of data storage and knowledge

In traditional storage systems, there is a central database which stores all the records. To prevent possible data losses in case of emergency, these systems can be set to do regular backups. However, a special care needs to be taken in order to ensure that no data was lost, damaged or tampered. Usually, the systems which guarantee all these properties are very expensive to afford and maintain.

In comparison, decentralized data platforms offer storage redundancy, but eliminate all security and maintenance risks. Additionally, the decentralized system can be configured in order to consume less storage if required [5].

1.4 Ownership of information

While in traditional data platforms data is typically owned by one party, decentralized platforms are highly focused on the problem of information ownership. Unlike traditional data platforms, they are designed to not to concentrate all stored information, as well as computing power on a single party or even in hands of some group.

But this doesn't mean that the information is not accessible nor is accessible to all parties. If we take blockchain as an example, every network participant in a decentralized system can obtain a copy of all information available, but the information itself can be both open or encrypted. However, in a properly set up decentralized system, one can be sure that the information recorded to the decentralized storage is permanent, and all data that belong to one party won't change its hands to another, which is guaranteed by data immutability property of such systems [6].

2 Decentralized Security Approaches

In order to protect a typical centralized computer system from all possible intrusions and attacks, which also applies to decision support systems, a lot of work must be performed. This, in turn, does not exclude the possibility of a hack and cannot guarantee that the system is fully protected. In other words, no matter how the system is protected, there is always a practical way to attack it.

Decentralized platforms, with blockchain technology as a primary example, offer a solution to above problems: they eliminate risks of tampering with the system and provide a data platform with a strictly determined way of how different parties can interact with the platform and rules of any possible data manipulations within the platform. Also, they offer a highly failure-resistant system, which means that even if half of all network nodes will crash, the system will still continue to operate normally [5].

Decentralized data platforms like blockchain guarantee by design that no single party can interrupt the normal, predictable functioning of the platform, as well as no data can be changed in an undefined way. In other words, there is no practical way to make a decentralized program produce a different result from expected, as well as tamper with any historical data [6].

2.1 Existing Decentralized Data Platforms

The first successful application of a public decentralized platform was released in 2008 by Satoshi Nakamoto, the name used by the unknown person or a group of people who originally developed Bitcoin – a peer-to-peer electronic cash system, which is still under a high demand today [9]. At that time it was barely possible to call Bitcoin a "decentralized platform", as the only functionality it served is virtual value transfers between parties. The term "decentralized platform" became well-recognized almost 7 years later, in 2015, with the release of Ethereum – world's first practically successful decentralized applications platform, which allowed people not only to transfer value over the internet but also to run any programs and computations which cannot be tampered with [10].

However, there are other emerging technologies that try to address many blockchain issues. We will briefly describe them below.

Blockchain. Bitcoin and Ethereum, as the most widely adopted decentralized platforms, are running using blockchain technology, which is one that is practically proven to be secure and stable. Blockchain's main disadvantage is its scalability issues, which other decentralized data platforms try to address. However, blockchain sharding, as the most recently introduced way to scale traditional blockchain platforms, offers a way to increase blockchain throughput without compromising security [8].

The name "Blockchain" appeared directly from an underlying algorithm meaning – the chain of blocks. Transactions in a blockchain are organized in blocks, and blocks reference each other. This structure with cryptographically secured references between blocks is called blockchain [8] (Fig. 2).

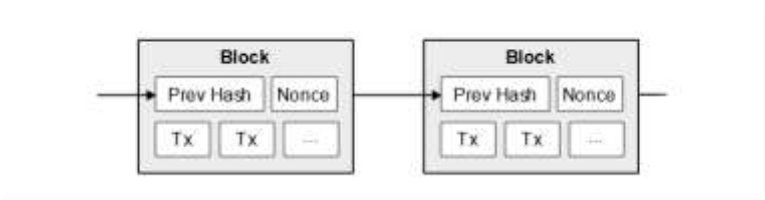


Fig. 2. Blockchain data structure

In blockchain, each next block must reference its previous block in order for the whole structure to be valid. This reference isn't just a link – it's a hash of a previous block. If a previous block's contents changes, its hash changes, and thus the reference to modified block becomes invalid. Because the whole network in blockchain validates blocks, every invalid block is just ignored, leaving no chances to someone to tamper with data and being unnoticed.

The block size in blockchain is intentionally limited, which leads to scalability issues [11]. If the block size wasn't limited, then fewer network participants could participate in forming a consensus because of hardware and network throughput limitations.

Hashgraph. Unlike blockchain, hashgraph does not pack transactions into blocks. Instead, transactions have cryptographic references to each other, forming a data structure which is also called hashgraph [12]. For example, in IOTA project, which is an early adopter of this approach, the global ledger forms a data structure called Tangle, which is a direct

acyclic graph in which each next transaction reference two previous [13]. The example of the hashgraph structure is depicted in figure 3.

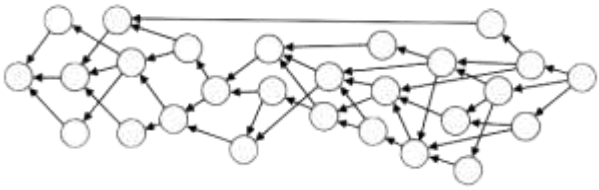


Fig. 3. Hashgraph (direct acyclic graph) structure

Hashgraph uses a probabilistic-based consensus algorithm, as an opposite to blockchain's deterministic consensus algorithms. For instance, to ensure that the particular transaction happened in hashgraph, a network participant uses a gossip protocol; it asks some of its peers about whether or not they have the same transaction recorded, and the transaction is considered valid when, for example, 2/3 of the peers respond positively.

Unlike blockchain, hashgraph doesn't have a limit of transactions. Its tangle can theoretically become very big and still be able to handle all incoming transactions. But, on the other side, hashgraph is less resistant to attacks. By having enough power, the continuous attack on a hashgraph ledger can create a "parasitic" hashgraph, which can prevent new transactions from happening as they might be considered invalid by more than 1/3 of the network, for example.

There are many examples where Ethereum and other similar technologies were adopted by businesses, government and healthcare, as well as many other fields of study. However, these platforms didn't get massively adopted yet, as they all suffer from scalability or security issues. There are many different higher-level solutions were introduced like sharding, lightning network, plasma, and others, but still, they are all vulnerable in case of weak algorithms used.

2.2 Existing Consensus Algorithms

A consensus algorithm in a decentralized network is a process used to achieve agreement on a single data value or a network state among distributed network participants. In Blockchain, a consensus algorithm is used to agree on the latest state of a ledger, which is supported by all network participants [14].

There are many consensus algorithms in existence, but there are just a few which are practically proven to work and are widely used as of 2018: Proof of Work, Proof of Stake and Delegated Proof of Stake. We will briefly introduce the main idea behind these algorithms and compare them to each other.

Proof of Work. In a Proof of Work consensus algorithm, the blockchain network agrees on a rule to accept the valid chain with the highest total complexity [15]. Everyone in a network can try to increase the total chain complexity, by adding new blocks with transactions. Network participants are economically incentivized to add transactions of other network participants to blocks they produce, as after forging a valid block they collect all transaction fees plus a block finding reward.

Adding each new block to a chain requires finding such a hash of a block that maps to a number which is below the given threshold called complexity. Finding this hash is not a trivial task; this task doesn't have a known solution due to the pre-image resistance property of a hash function used. Thus, the only way to find a valid hash of the block is to try all possible values of a random number nonce which is used in hashing.

The first network participant called "miner" who generates a valid hash from a block broadcasts it to a network, while all other network participants accept this block because its hash is below the given complexity for a block. The next block's complexity is deterministically computed from the complexity of previous blocks and time required for its mining so that no matter how powerful computers in the network are, there is always a stable approximate time between new blocks added to the network (15 seconds for Ethereum).

Because finding a right hash for a block is time and computational resources consuming task, network attracts more and more incentivized parties that try to find a right hash, making network complexity enormously big, which, in turn, makes any possible attacks to this network impractical.

Proof of Stake. Proof of Work algorithm requires a lot of energy for mining [14, 16], and this creates two problems:

- A lot of energy is consumed for a very little amount of useful work (performing transaction in a blockchain).
- A risk that parties which have supercomputers and computational data centers can break the network by addressing all their computational capabilities to it.

Proof of stake algorithm solves this problem by introducing virtual mining – a pseudo-random deterministic way of identifying which party can produce the next block, based on numerous factors. The primary factor which Proof of Stake takes into account is the party's stake – the number of virtual currency the party owns. In Ethereum, for example, this currency is Ether. This means the more currency the party has, the more blocks it can produce [17].

Proof of Stake doesn't solve the problem of one party owning almost all currency, as well as Proof of Work doesn't solve the problem of one party owning all computational powers. However, it is more efficient as it requires a very few computing resources when compared to Proof of Work.

Delegated Proof of Stake is one of the most complex consensus algorithms widely used today. It works similar to Proof of Stake with the difference in the next block producer selection algorithm. Delegated Proof of Stake block producer mechanism works more like an election, where all network participants can vote for block producers using their in-network currency, thus, increasing their chances to become actual block producers [18].

Unlike Proof of Stake, there isa limited number of block producer involved, which allows higher scalability comparing to Proof of Work and Proof of Stake, as only a limited number of network participants do actual computations, while others just consume them as is.

2.3 Typical Access Control Models

Because there are tasks that one particular decentralized data platform cannot handle, such as heavy computations or high throughput and reliability at the same time, there are several different approaches of how the decentralized network can be set up in order to fit the needs [5].

Public Decentralized Data Platforms. Public decentralized data platform is a decentralized data platform intended for a public use, meaning that any party can transact in its network. Just like the internet, but immutable, predictable and censorship-resistant. Because the network of a decentralized data platform is public, any network participant can transparently see and verify each transaction in it, ensuring that the data was not corrupted or changed [19, 20].

Ethereum is one of the most well-known public decentralized data platforms, which is able to handle a maximum of 15 transactions per second. Handling more transactions per second makes the system more

centralized, as more resourced is required to support the network, while not every network participant can afford to have this resources available.

Public decentralized data platforms like Ethereum allow providing a maximal trust to a software solution based on them. However, sometimes, these platforms are not scalable enough, and for some cases, there is a point in using private or permissioned platforms.

Private Decentralized Data Platforms. Because public decentralized data platforms cannot afford big throughput due to security reasons, there is an option to launch a private network, which will be only accessible to a particular number of parties [20]. Making the decentralized network private means that it eventually loses one of its properties – decentralization, because it is being controlled by one party which keeps it private. However, several approaches could help to preserve trust in the system:

- Occasionally publishing a network state on a regular basis to another decentralized public ledger, which can be used to validate the state of a private network in the future.
- Allowing external provisioners to validate the network state on-demand, by introducing a public registry of backups made by a system regularly.

While private networks are closer to traditional centralized platforms, they borrow some useful properties from decentralized ones, like immutable network history and fault tolerance.

Permissioned Decentralized Data Platforms combine the properties of public and private decentralized data platforms, finding the golden mean between two. These platforms are often referenced as consortium platforms (for example, consortium blockchain). They function just like public platforms but have more restrictions on who can access the platform and, more importantly, who can perform which transactions [19, 20].

Speaking of the blockchain technology, in a permissioned decentralized system there may be a pre-defined set of parties who can assemble blocks. For instance, in a bank system where more than 10 banks are involved, each bank as a network participant can have an equal weight, leading to everyone have the same influence on the network. The rules are strictly defined in a system, and once one instance does not follow these rules it exists to the chain which is considered invalid by all other parties, thus leaving the network.

The biggest example of permissioned blockchain today is Hyperledger Fabric, providing a flexible blockchain-based platform that can be set up for almost any needs. While it is highly configurable, there is still no possible configuration in which the blockchain will be scalable, highly available and secure at the same time.

3 Security Model for High-Risk Systems

Decentralized data platform application can bring the following properties to the high-risk systems:

1. Data security and integrity.
2. Data immutability and transparent history of changes.
3. Reliable and fault-resistant storage.
4. Reliable and predictable data processing.

Regarding decision support systems, there are several possible applications. We describe 2 of them, based on a fully public access control model and semi-public permissioned model.

Private and consortium security models of decentralized data platforms are quite challenging to set up properly, so we do not recommend setting them up for individual organizations, rather than joining to already existing private or consortium decentralized data platforms.

3.1 System Security Model Based on Public Decentralized Data Platforms

One of the main applications is to build a system fully on top of a decentralized platform. This is always the most desirable scenario, however, it has two disadvantages:

1. It requires developing a solution which doesn't have functional bugs or special method development which allows updating the solution without further centralization risks.
2. The reliable public decentralized platform which solves a scalability problem does not exist yet, which limits the number of operations that can be performed in a decentralized ledger.

In a case of the full decentralization, regarding decision support system, experts are using the client which is connected to a decentralized data platform directly, making the direct contributions to a global public

ledger of a decentralized data platform. The entered data becomes immutable immediately after the entry, and all its possible future changes are practically excluded. The diagram demonstrating the process of information entry, processing and verifying is depicted in figure 4.

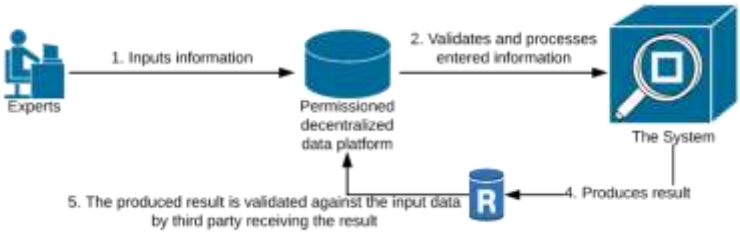


Fig. 4. System security model where experts directly submit information to a permissioned decentralized system.

The client used for the data entry by experts uses the decentralized identity – an account in terms of the decentralized data platform. These accounts are registered before the data entry

Thus, after the result is processed by the decision support system, each recommendation given by the system is compared and verified against input data, which is 100% legit. Moreover, in case of decentralized data platform can handle the load and computing resources required to process the input data and produce a decision, the whole decision support system can be built on top of the decentralized data platform.

3.2 System Security Model Based on Unscalable Public Decentralized Data Platforms

Taking into account that the public security model currently does not fit all the needs of a scalable and, as a result, reliable system, we suggest another approach to creating a reliable and secure trustless system for high-risk decision support systems based on permissioned decentralized data platforms.

This security model decreases the required space of stored data within the decentralized platform, making it cheaper to use. For example, it can reduce the cost of each complete decision support process to $0.03 in equivalent, when using Ethereum decentralized platform usage (the average cost of one transaction in the main Ethereum network, as of 11/8/2018).

In this security model, the data is entered into the system using a regular centralized database. After the data is entered by all experts, this data is cryptographically hashed with a strong hashing algorithm and is published to a public decentralized data platform as a proof of the data state. Later, the input information is disclosed and experts confirm that the resulting hash corresponds to an input data indeed. This data can be confirmed by any provisioning expert and saved for further proof that the output was produced from the legit input. This procedure is somehow similar to obtaining a digital signature for every information entered to the system, but without a practical way to change the signature itself.

Figure 5 demonstrates this approach. The only information that goes to a decentralized secure ledger is a hashed information, which stands as a proof of legit input data that is compared to the original input after the decision support process is finished.

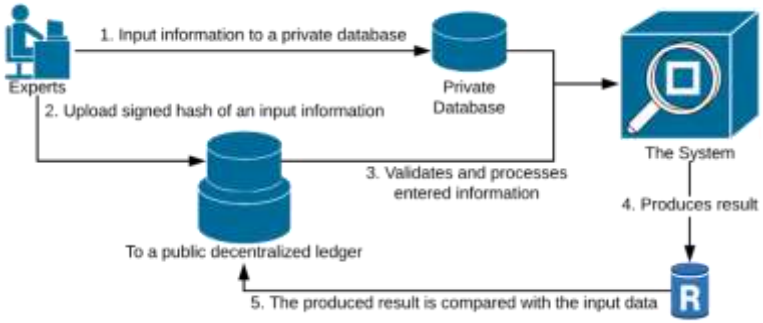


Fig. 5. System security model example where only hashed information is submitted to a decentralized data platform, making it cheaper to operate

In this scenario it is important to mention that the original data can be lost, resulting in having a hash that does not correspond to any data. But at least, for the consumer, this will mean that the process of decision support was not organized properly and the produced decision can be forged.

Conclusion

High-risk systems, including decision support systems, require the most secure program and infrastructure environment to function. Decentralized data platforms, which is an emerging technology nowadays, is the only way for systems and applications to *practically* exclude any tampering

risks and intrusion possibility. By utilizing the decentralized security model proposed in this article high-risk systems can expect their data and processing to be safe and predictable regardless of how the internal system is developed and maintained.

References

1. Saaty, T. L. (2010). Principia Mathematica Decernendi - Mathematical principles of decision making - Generalization of the Analytic Network Process to neural firing and synthesis. Pittsburg: RWS Publications.
2. Tsyganok V., Kadenko S., Andriychuk O., Roik P. Usage of multicriteria decision-making support arsenal for strategic planning in environmental protection sphere / *Journal of Multi-Criteria Decision Analysis.* 2017;**24**:227–238.
3. Driscoll, J.W., 1978. Trust and participation in organizational decision making as predictors of satisfaction. Academy of management journal, 21(1), pp.44-56.
4. Tsyganok V.V., Kadenko S.V. & Andriichuk O.V. Using different pair-wise comparison scales for developing industrial strategies. *International Journal of Management and Decsion Making*. – 2015. – vol. 14, issue 3. – P. 224-250.
5. Tsyganok V.V.,KadenkoS.V., Andriichuk O.V. Usage of Scales with Different Number of Grades for Pair Comparisons in Decision Support Systems / *International Journal of the Analytic Hierarchy Process*. – 2016. – vol.8, issue 1. – P.112-130.
6. Андрійчук О.В. Метод змістової ідентифікації об'єктів баз знань систем підтримки прийняття рішень / *Реєстрація, зберігання і обробкаданих.* – 2014, - Т.16, №1 – С.65-78.
7. Dhillon, G. and Torkzadeh, G., 2006. Value- focused assessment of information system security in organizations. Information Systems Journal, 16(3), pp.293-314.
8. Pilkington, M., 2016. 11 Blockchain technology: principles and applications. Research handbook on digital transformations, p.225.
9. Puthal, D., Malik, N., Mohanty, S.P., Kougianos, E. and Yang, C., 2018. The blockchain as a decentralized security framework. IEEE Consum. Electron. Mag., 7(2), pp.18-21.
10. Seebacher, S. and Schüritz, R., 2017, May. Blockchain technology as an enabler of service systems: A structured literature review. In International Conference on Exploring Services Science (pp. 12-23). Springer, Cham.
11. Swan, M., 2015. Blockchain: Blueprint for a new economy. "O'Reilly Media, Inc.".
12. Nakamoto, S., 2008. Bitcoin: A peer-to-peer electronic cash system.

13. Wood, G., 2014. Ethereum: A secure decentralisedgeneralised transaction ledger. Ethereum project yellow paper, 151, pp.1-32.
14. Croman, K., Decker, C., Eyal, I., Gencer, A.E., Juels, A., Kosba, A., Miller, A., Saxena, P., Shi, E., Sirer, E.G. and Song, D., 2016, February. On scaling decentralized blockchains. In International Conference on Financial Cryptography and Data Security (pp. 106-125). Springer, Berlin, Heidelberg.
15. Hoxha, L., 2018. Hashgraph the Future of Decentralized Technology and the End of Blockchain. European Journal of Formal Sciences and Engineering, 1(2), pp.29-32.
16. Popov, S., 2016. The tangle Whitepaper. cit. on, p.131.
17. Zheng, Z., Xie, S., Dai, H., Chen, X. and Wang, H., 2017, June. An overview of blockchain technology: Architecture, consensus, and future trends. In Big Data (BigData Congress), 2017 IEEE International Congress on (pp. 557-564). IEEE.
18. Vukolić, M., 2015, October. The quest for scalable blockchain fabric: Proof-of-work vs. BFT replication. In International Workshop on Open Problems in Network Security (pp. 112-125). Springer, Cham.
19. O'Dwyer, K.J. and Malone, D., 2014. Bitcoin mining and its energy footprint.
20. Bentov, I., Gabizon, A. and Mizrahi, A., 2016, February. Cryptocurrencies without proof of work. In International Conference on Financial Cryptography and Data Security (pp. 142-157). Springer, Berlin, Heidelberg.
21. Larimer, D., 2014. Delegated Proof of Stake. Bitshares. org. From Bitshares.org, last accessed 2016/11/21.
22. Cachin, C., 2016, July. Architecture of the hyperledger blockchain fabric. In Workshop on Distributed Cryptocurrencies and Consensus Ledgers (Vol. 310).
23. Zheng, Z., Xie, S., Dai, H.N. and Wang, H., 2016. Blockchain challenges and opportunities: A survey. Work Pap.–2016.
24. LNCS Homepage, http://www.springer.com/lncs, last accessed 2016/11/21.

ВИЗНАЧЕННЯ ЕКСПЕРТНИХ ГРУП ДЛЯ НАУКОВОЇ ЕКСПЕРТИЗИ

Балагура І.В.[1] , Ланде Д.В. [1], Андрущенко В.Б.[2], Горбов І.В. [1]
[1] Інститут проблем реєстрації інформації Національної академії наук України, м.Київ
[2] Національний центр співробітництва із ЄС у сфері науки і технологій, м.Київ
balaguraira@gmail.com

На сьогодні не існує уніфікованої моделі для вибору експертів або фахівців за формальними ознаками, тому в даному проекті запропоновано підхід до розв'язання задачі пошуку експертів на основі мереж співавторів. Запропоновано метод пошуку наукових колективів, що дозволяє за заданими термінами певної предметної області на основі мереж термінів знаходити групи вчених, діяльність яких відповідає тематиці. Роботу алгоритму реалізовано на основі інформації з Web of science, проведено порівняння роботи алгоритму із системою Aminer. Запропонований метод дозволяє проводити пошук за більш гнучкими запитами. Доцільно використовувати бази даних авторів підчас пошуку наукових колективів для виключення однакових прізвищ та отримання детальної інформації.

Keywords: *мережі співавторів, мережі термінів, експерти, наукові бази даних, наукові публікації.*

Вступ

На сьогодні не існує уніфікованої моделі для вибору експертів або фахівців за формальними ознаками, тому в даному проекті запропоновано підхід до розв'язання задачі пошуку експертів на основі мереж співавторів. Актуальним завданням є підбір компетентних експертів, залучення науковців до вирішення важливих державних завдань. Крім того, якісна експертиза необхідна і для реформування самої науки, що розпочато керівництвом держави з прийняттям закону Україні "Про наукову та науково-технічну діяльність" від 26 грудня 2015 року [1]. Актуальність пошуку та формування експертних колективів пов'язано також з участю України в програмі "Горизонт-2020" та українських науковців в якості експертів конкурсних проектів [2].

Визначення пріоритетних напрямів, за якими Україна зможе гідно представити наукові розробки світовій спільноті та посісти гідні місця за результатами експертної оцінки проектів, на сьогодні є одним з пріоритетних завдань держави. Формування експертних колективів є необхідним для виконання цього завдання. Запропонований метод може виступити важливим інструментом для об'єктивного прийняття рішення при формуванні подібних колективів. Використання запропонованого методу дозволить проводити об'єктивний пошук міжнародних експертів для завдань оцінки та реформування окремих галузей в державі на основі відповідних наукових праць. Розв'язок задачі швидкого і кваліфікованого визначення експертних груп сприятиме ефективному виконанню Закону України «Про наукову і науково-технічну експертизу» та Закону України «про інноваційну діяльність» [3].

Мета дослідження полягає у розробці методу та його апробації для кваліфікованого і об'єктивного визначення експертних груп фахівців за ознаками володіння спеціальними знаннями для здійснення наукової чи науково-технічної експертизи.

1 Існуючі підходи пошуку колективів експертів

На багатьох українських і закордонних веб-сайтах наукових, культурних та інших можна знайти оголошення із пошуку експертів. Це метод анкетування, його проводять в очній або заочній формі [4]. Для кандидатів висуваються певні вимоги з досвіду, навичок роботи та посади для виконання експертизи в тій чи іншій галузі. Даний метод пошуку експертів є на даний момент найпоширенішим у світі. Проте не для всіх видів робіт можна оголошувати тривалий відкритий конкурс робіт та не в кожному випадку спеціаліст зверне увагу на оголошення. Тому пошук експертів шляхом анкетування та проведення конкурсів не є найбільш ефективним.

Другим способом пошуку експертів є використання соціальних мереж з профілями спеціалістів таких як ResearchGate (https://www.researchgate.net/), LinkedIn (https://www.linkedin.com/) [5]. Даний метод подібно до пошуку експертів методами анкетування дозволяє вивчити інформацію про спеціаліста на основі відповідного профілю, проте в даному випадку не витрачається час на проведення конкурсу, а також існує можливість розглянути кандидатури експертів незалежно ві країни проживання та власної зацікавленості.

Для наукової експертизи проведення пошуку експертів можливе на основі наукометричних баз даних, що містять профілі науковців з інформацією про публікації, цитування та інші науко метричні показники. Наукові профілі можна знайти в Google scholar, Scopus, Web of science та інших базах даних. Також існують ресурси для об'єднання інформації про науковців з різних баз даних, серед них: ORCID(https://orcid.org/), «Бібліометрика української науки» (http://www.nbuviap.gov.ua/bpnu/), «Науковці України»(http://irbis-nbuv.gov.ua), AMiner (https://aminer.org/) та інші. ORCID – всесвітньо відомий сервіс, що забезпечує можливість об'єднати та оцінити публікаційну активність науковця та визначити особистість відповідно до ідентифікатору. «Бібліометрика української науки», «Науковці України» створені національною бібліотекою України ім. В.І. Вернадського об'єднання інформації у першому випадку наукових профілів із світових наукометричних ресурсів, у другом – інформації із авторефератів дисертацій. AMiner – база даних науковців із штучного інтелекту, що містить 130614292 науковців створена у Китайській народній республіці. AMiner містить детальну інформацію про автора: розподіл публікацій за тематиками, цитування, індекс Гірша, g-індекс, мережу співавторів, місце роботи, основні навички та визначає ранг автора серед інших відповідно до одного із критеріїв за певним науковим напрямом. В AMiner проводити сортування можна відповідно до країни, мови публікацій, статі, ресурсів про авторів, індексу Гірша та по детальним рубрикам, а також за визначеним концептом [6]. Основним недоліком таких сайтів є обмеженість пошуку відповідно до визначених тематичних рубрик.

В роботі [7] проведений порівняльний аналіз методів автоматичного пошуку експертів. Авторами моделі пошуку експертів поділені на імовірнісні (розрахунок для авторів або тематик), ранжування та мережеві (PageRank, HITS). Вибір методу залежить від мети пошуку експертів та джерела даних. Саме вхідні дані в найбільшій мірі визначають результат, тому на першому етапі аналізу найбільш важливо обрати базу даних публікацій.

2 Метод пошуку колективів експертів

Для пошуку експертів вирішення державних завдань доцільно проводити пошук серед наукових колективів, що сформовані в межах певних наукових шкіл та обирати спеціалістів з різних колективів. Джерелом для визначення експертів мають бути

авторитетні наукові бази даних, що містять інформацію про науковців із України. Запропоновано використовувати аналіз мереж співавторів, зокрема для визначення наукових колективів – алгоритми на основі модулярності, що реалізовані в більшості автоматизованих пакетів для обробки складних мереж та довели свою ефективність на багатьох прикладах [8]. Для побудови мереж термінів запропоновано використовувати методи, представлений у [9].

Таким чином запропоновано метод пошуку експертів, що містить наступні етапи:

1.Визначається галузь, в якій проводиться пошук та відповідні концепти.

2.Фільтруються дані із наукової бази даних.

3. Будується мережа співавторів, обчислюються основні характеристики мережі, проводиться розбиття мережі на колективи на основі модулярності.

4. Будуються мережі термінів для кожного колективу співавторів.

5. Будуються окремі гетерогенні мережі, що поєднують співавторів та терміни відповідно до сформованих у 3 кластерів.

6. Визначаються мережі із найбільшою кількістю відповідних концептів.

7.Проводиться ранжування науковців в межах окремих колективів за показниками центральності відповідно до [10].

8. Формується список експертів відповідно.

Для можливості порівняти результати визначені з допомогою запропонованого методу із результатами бази даних Aminer проведено аналіз масиву бази даних Web of science в галузі штучного інтелекту за концептом link prediction для науковців, що працюють Китайській народній республіці. У вибірці отримано 674 науковці. Мережа співавторів представлена на рис.1 (за допомогою VOSviewer)

В результаті отримано всього п’ять наукових колективів, що працюють над заданою тематикою. В базі даних Aminer знайдено 66 профілів за відповідним концептом. В кожному профілі автора вказано відсоток праць із певної тематики, проте немає можливості простежити вживання в роботах окремих концептів. В запропонованому методі на наступному етапі будуються мережі термінів (рис.2). Поєднання мереж співавторів та мереж термінів дає можливість визначити колективи, що найбільше працюють за заданою тематикою. На наступному етапі можна провести ранжування та виділити лідерів наукових напрямів. Бази даних

науковців обов'язково необхідно використовувати після отриманих на основі мереж співавторів та термінів. Такі бази даних дають можливість визначити досвід доробок авторів, відокремити авторів з однаковими прізвищами. Представлений метод пошуку експертів потребує більш детальної апробації для українських науковців на основі баз даних наукових публікацій.

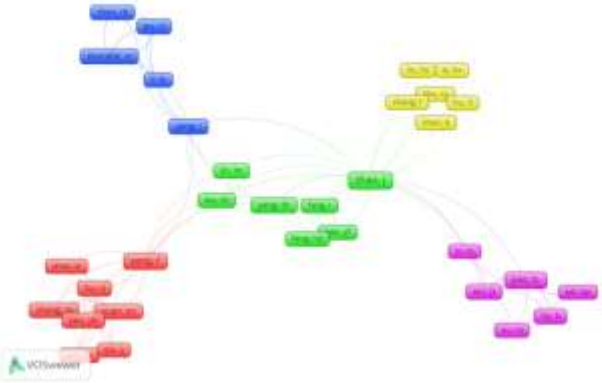


Рис.7. Мережа співавторів

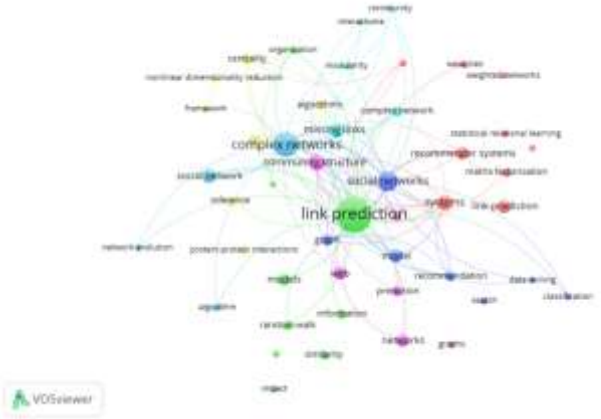


Рис.2. Мережа термінів

Висновки

Запропоновано метод пошуку наукових колективів, що дозволяє за заданими термінами певної предметної області на основі мереж термінів знаходити групи вчених, діяльність яких відповідає тематиці. Показано, що запропонований метод дозволяє проводити пошук за більш гнучкими запитами в порівнянні із існуючими базами даних експертів. Доцільно використовувати бази даних науковців підчас пошуку наукових колективів для виключення однакових прізвищ та отримання детальної інформації про експерта.

Публікація містить результати досліджень, проведених за грантом Президента України за конкурсним проектом Ф75/173-2018 Державного фонду фундаментальних досліджень.

Посилання

1. Закон України "Про наукову та науково-технічну діяльніст" [електронний ресурс] –Режим доступу: http://nucpi.nas.gov.ua/news/item/67-zakon-ukraini-pro-vischu-osvitu.html
2. Горизонт 2020. Національний портал [електронний ресурс] – Режим доступу: https://h2020.com.ua/
3. Закон України «Про наукову і науково-технічну експертизу» [електронний ресурс] –Режим доступу: http://zakon.rada.gov.ua/laws/show/51/95-%D0%B2%D1%80
4. Рогушина Ю.В. Використання організаційних онтологій для пошуку експертів у нових предметних областях /Ю.В. Рогушина, А.Я.Гладун//Проблеми програмування.-2007.-с.73-83
5. C. Wei, W. Lin, H. Chen, W. An, and W. Yeh. Finding experts in online forums for enhancing knowledge sharing and accessibility, Computers in Human Behavior, vol.51, pp.325-335 (2015)
6. Robin Brochier, Adrien Guille, Benjamin Rothan, Julien Velci Impact of the Query Set on the Evaluation of Expert Finding Systems, CoRR, abs/1806.10813 (2018)
7. Lin, S., Hong, W., Wang, D., & Li, T. A survey on expert finding techniques. Journal of Intelligent Information Systems, 49(2), 255–279.(2017)
8. Балагура І.В., Ланде Д.В. Лінгвістичні дослідження взаємозв'язків науковців на основі аналізу реферативної бази даних «Україніка наукова», Реєстрація, зберігання і обробка даних. Т.16, №3 –С.45-53.(2014)
9. Dmitry Lande, Andrey Snarskii, Elena Yagunova The Use Of Horizontal Visibility Graphs To Identify The Words That Define The

Information Structure Of The Text CEUR Workshop Proceedings. Vol-1108 urn:nbn:de:0074-1108-1. ISSN 1613-0073. Selected Papers of the 15th All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections" Yaroslavl, Russia, October 14-17, 2013. - P. 158-164.(2013)

10. Горбов І.В., Каденко С.В., Балагура І.В., Манько Д.Ю., Андрійчук О.В. Визначення потенційних експертних груп науковців в мережі співавторства з використанням методів підтримки прийняття рішень, Реєстрація, зберігання і обробка даних.,Т.15, №4 –С.87-97. (2013)

MATHEMATICAL MODELS OF CLOUD COMPUTING WITH ABSOLUTE - RELATIVE PRIORITIES OF PROVIDING COMPUTER RESOURCES TO USERS IN CONDITIONS OF FUNCTIONING FEATURES AND FAILURES

Matov A.Y.
Institute for information recording of NAS of Ukraine, Kyiv,Ukraine

Analytical models of cloud systems (CS) are developed as queuing systems with a mixed discipline of resource allocation. The models take into account failures and various functioning features and have arbitrary distribution laws for many stochastic processes. Such models for the CS are used for the first time. One of the main indicators of the effectiveness of the CS are indicators based on the evaluation of the time characteristics of these systems. Violation of the permissible time constraints, for example, the response time of the cloud system, adversely affects the efficiency of solving the target tasks of the user. This is particularly important for real-time systems and, first of all, for specific information systems built using private cloud systems.

General description of the models is as follows. The input of the cloud system, which implements a mixed service discipline (with relative and absolute priorities), receives N Poisson flows of requests for resources with corresponding N priorities. The duration of application processing of various flows have their own arbitrary distribution laws. An application with relative priority interrupted by applications with absolute priority, returns to the queue. Two disciplines of the resumption of service A and B are considered. Within the same priority, applications are processed on a first-come, first-served basis.

The CS fails according to the Poisson law, and is restored under an arbitrary law. During the recovery period, elements of adaptation to failures are used: applications of some flows to the queue are accepted and accumulated, while others are not accepted (the discipline of the queue replenishment, I and II, respectively,). The denial of service device can occur both during its free state and during the application processing. Two disciplines of resumption of service after restoration C and D are considered. An interrupted application is processed from the point of its interruption. The combination of disciplines - resumption of service and queue replenishment - allows us to consider

independent models of various types of systems that have the respective designation.

Various functioning features include various combinations of disciplines A, B, C, D, I and II.

Keywords: *cloud computing, discipline of providing computing resources, adaptation and optimization of service disciplines, efficiency of adaptation, mixed service discipline, mathematical model.*

Introduction

The development of mathematical models of cloud computing or information systems created using clouds is an important area for identifying and improving their characteristics [1,3,10…12]. Cloud computing (CC) is an object with a high level of uncertainty in the functioning process, the main factors of which are [1,3]:

-problem of the flow of requests for computing resources (CR);

-presence of the required CR and the accidental time of their use by customers;

-accidently failure of the infrastructure of the CC and the time of their elimination;

-necessity to provide certain time characteristics for a number of clients, for example, the response time of the CC;

-necessity of optimal use of PR depending on the cost of time delay of customers ordered results of calculations and operating conditions;

-necessity of introduction of adaptation in the process of functioning CC in order to provide certain time characteristics for a number of clients and optimal use OP.

One of the main indicators of the effectiveness of CC is the indicators based on the assessment of the time characteristics of these systems. Violation of permissible time constraints, for example, the response time of the CR, affects the effectiveness of the solution of user targets, which is of particular importance for real-time systems. First of all it concerns special information systems, which are built using private CC.

The stochastic nature of the main factors and the necessity of quantification of mass processes on the basis of the theory of probability determines the use of the theory of mass service. Then it is possible and appropriate to use the technology of the dynamic adaptive mixed

discipline of providing CR (maintenance) to users of the CC [1] as mechanisms of adaptation of the CS.

Analytical models for calculation of time characteristics are offered in the conditions of the features of the functioning of the CC using a mixed discipline of service with absolutely relative priorities and taking into account failures. Models are based on works [2, 4 7, 9].

Description of the model of cloud infrastructure operation with mixed service discipline and adaptation to failures

Let the input of the CC system, in which the discipline of service with a relatively absolute priority is implemented, arrive N Poisson flows of applications of intensity $\lambda(m,n)$ $(m=\overline{1,M}$, $n=\overline{1,N_m})$. These flows are aligned with N priorities [2].

The duration of the maintenance of applications of priority (m, n) is a random variable with a distribution function $B_{m,\ n}(t)$, the first *b (m, n)* and the second $b^{(2)}(m,n)$ start point.

An application of priority (m, n) whose service is interrupted by applications from groups with $\overline{1,m-1}$, numbers is returned to the queue. Updating its service is possible either after servicing all interrupted applications (maintenance discipline A), or after servicing all interrupted applications and all applications for accumulated flows, the *m* group with $\overline{(m,1),(m,n-1)}$ numbers (discipline of service upgrade B) .

The serving device (CC) fails in accordance with the Poisson law with the λ_0 parameter. The period of recovery of the device is a random variable that has an arbitrary distribution law *$B_o(t)$* with the first b_0 and second b_0^2 initial moments.

During the restoration of the service device, requests of some streams in the queue are accepted, while others are not accepted. This condition is given by the matrix-row of coefficients n_i, $i=\overline{1,N}$, , and in the case if requests of the $n_i=1$ stream are accepted in the queue, and if requests $n_i=0$ are denied.

Adaptation to bounce will be that in the period of recovery device incoming applications can either accumulate in the queue (discipline replenishment queue I), or receive a refusal and leave the system (discipline replenishment queue II).

Failure of the servicing device can occur both during its free state and during service of the application. In the latter case, the renewal of the service is carried out either from the interrupted application, if there are no applications interrupting its service, (the discipline of the renewal of service C), or from applications of the senior relative priority of the corresponding group, if any (discipline of renewal of service D).

In case of repeated receipt of the servicing device, the interrupted application shall be maintained from the place where it was interrupted. Within one priority, applications are served in the order of receipt.

The combination of service updating disciplines and queue replenishment allows you to consider independent models of different types of systems that have the proper designation. Different features of functioning consist of various combinations of disciplines A, B, C, D, I and II.

Let CC be in stationary mode, which $R_M < K_r$ condition is for systems of type I, and for systems of type II - $R_M < 1$. Here $R_M = \sum_{m=1}^{M} \sum_{n=1}^{N} \rho(m,n)$ - total loading of the device applications ($(\rho(m,n) = \lambda(m,n) b(m,n)$ - loading of the device (m, n) - applications), and $K_r = 1/(1+\rho_0)$ - the system readiness coefficient ($(\rho_0 = \lambda_0 b_0$ - loading the device with refusals).

It is necessary to determine the average $v(m,n)$ time spent in the system of applications of each (m, n) -priority, ie, the response time of the system CC.

Definition of time characteristics of a model of a system of type AS-I.

To determine the average time of applications in the system (time response systems) type AS-I use the known direct method [2].

Let some application (*j, k*) be a priority in the system. The average duration of this application in the system *v* (*j, k*) consists of the average waiting time in the queue *w* (j, k) and the average service time b (*j, k*):

$$v(j,k) = w(j,k) + b(j,k) \qquad (1)$$

The average waiting time in the queue w (*j, k*) consists of the average waiting time before service and the average standby time in the interrupted state u (*j, k*):

$$w(j,k) = w_H(j,k) + u(j,k). \qquad (2)$$

The last term in this formula is due to the interruptions in the maintenance of the application (j, k) -priority of applications from groups $\overline{1, j-1}$ and denials, that is:

$$u(j,k) = u_3(j,k) + u_0(j,k). \qquad (3)$$

Average time from the beginning of service (j, k) - application to completion is the average full time of service:

$$\Theta(j,k) = b(j,k) + u(j,k). \qquad (4)$$

Let's start with the calculation u (j, k), for which we apply the approach described in [2].

During the service (j, k) -supply on average will occur $b(j,k)\Lambda_{j-1}$ interruptions where $\Lambda_{j-1} = \sum_{m=1}^{j-1}\sum_{n=1}^{N_m}\lambda(m,n)$ the intensity of the total flow of interrupted applications.

As a result of these interruptions (j, k), the application returns to the queue and waits for the termination of service interruptions that will continue in $b(j,k)R_{j-1}$ average units of time

where $R_{j-1} = \sum_{m=1}^{j-1}\sum_{n=1}^{N_m}\lambda(m,n)b(m,n)$. (5)

During this time, applications from groups $\overline{1, j-1}$ will be received, which will lead to an increase in waiting time (j, k) - applications for value $b(j,k)R_{j-1}^2$. In addition, the service of these applications will be accompanied by additional accumulation of applications of the same priorities, requiring service before (j, k) -payment. This process is endless, with supplements to the waiting time (j, k) -positions form a declining geometric progression with a denominator $R_{j-1} < 1$. The sum of members of such geometric progression is the mean time of all service interruptions (j, k) -request:

$$T^{(1)} = b(j,k)\frac{R_{j-1}}{1-R_{j-1}}. \qquad (6)$$

In the mean time $T^{(1)}$, the device will fail $T^{(1)}\lambda_0$, resulting in it will be restored within $T^{(1)}\lambda_0 b_0 = T^{(1)}\rho_0$ units of time. Since in the system type

AS-I during the period of recovery the device again receives applications that continue to accumulate in the queue, then after the device is restored, the average waiting time (*j, k*) -supply in the interrupted state will increase by

$$\mathrm{T}^{(2)} = \mathrm{T}^{(1)}\rho_0 \frac{R_{j-1}}{1-R_{j-1}} = b(j,k)\rho_0 \frac{R_{j-1}^2}{(1-R_{j-1})^2}. \quad (7)$$

During this time there may be a refusal of the device, the restoration of which will be accompanied by the accumulation of new applications served before (*j, k*) -payments, etc.

The total time of all applications service interruptions (*j, k*) -priority of $\overline{1, j-1}$ application groups, taking into account device refusals $u_з(j,k) = T^{(1)} + T^{(2)} + \ldots + T^{(\infty)}$. This expression represents the sum of two infinitely decreasing geometric progressions. After calculating the sum of the members of each of them and compiling the results, we get:

$$u_з(j,k) = b(j,k)\frac{R_{j-1}}{K_r - R_{j-1}}. \quad (8)$$

Similarly, the average waiting time (j, k) is determined in the interrupted state due to device refusals $u_0(j,k)$. The only difference is the beginning of reasoning. During the service (*j, k*) -supply, the device will fail on $b(j,k)\lambda_0$ average, which will result in its restoration within $b(j,k)\rho_0$ units of time. Taking into account the possibility of accumulation in the period of device renewal and priority service of applications with absolute priority from $\overline{1, j-1}$ group, the average waiting time (*j, k*) -payments will increase by

$$b(j,k)\rho_0 \frac{R_{j-1}}{1-R_{j-1}}.$$

During this time, the device can again be denied, which additionally increases the waiting time (*j, k*) - request for value $b(j,k)\rho_0^2 \frac{R_{j-1}}{1-R_{j-1}}$ etc.

In the final analysis, we get:

$$u_0(j,k) = b(j,k)\frac{K_r\rho_0}{K_r - R_{j-1}}. \quad (9)$$

Then the total average waiting time (*j*, *k*) -request in the interrupted state:

$$u(j,k) = b(j,k)\frac{R_{j-1} + K_r\rho_0}{K_r - R_{j-1}}\ , \qquad (10)$$

and the total average service time (*j*, *k*) -request:

$$\Theta(j,k) = b(j,k)\frac{1}{K_r - R_{j-1}}\ . \qquad (11)$$

Now calculate $w_H(j,k)$. Before (*j*, *k*) -request entered the system for the first time, the following should be done:

1) the device is restored

2) an application has been served from $\overline{1,j}$ or groups of submissions of the served application from the $\overline{j+1,M}$ groups;

3) service requests from $\overline{2,j}$ groups interrupted by applications from $\overline{1,j-1}$ groups;

4) service requests from $\overline{1,j}$ groups interrupted by denials of the device;

5) existing requests for streams with numbers $\overline{(1,1),(j,k)}$ are served;

6) service requests flowed with numbers $\overline{(1,1),(j,k-1)}$ received during the waiting time (*j*, *k*) -request, taking into account device refusals.

For the average duration of these events, we write the equation:

$$w_H(j,k) = \sigma_0 + \sigma(j,k) + \eta(j,k) + \eta_0(j,k) +$$
$$+\sum_{m=1}^{j-1}\sum_{n=1}^{N_m} w_H(m,n)\rho(m,n) + \sum_{n=1}^{k} w_H(j,n)\rho(j,n) + \quad . \qquad (12)$$
$$+[\sigma_0 + z_H(j,k)]\frac{R_{j,k-1}}{K_r - R_{j,k-1}} + z_H(j,k)\frac{K_r\rho_0}{K_r - R_{j,k-1}}$$

Here

$\sigma_0 = K_r\rho_0\Delta_0$ - average time for updating the device in the presence (*j*, *k*)-position: $K_r\rho_0$ - probability of recovery of the device [2], $\Delta_0 = b_0^{(2)} / 2b_0$;

$\sigma(j,k) = \sum_{m=1}^{j} \sum_{n+1}^{N_m} \rho(m,n)\Delta(m,n)$ – average time for the maintenance of the application by the device in the presence (j, k)-request: $\Delta(m,n) = b^{(2)}(m,n) / 2b(m,n)$;

$\eta(j,k) = \sum_{m=2}^{j} \sum_{n=1}^{N_m} \frac{R_{m-1}}{K_r - R_{m-1}} \rho(m,n)\Delta(m,n)$ - average time to receive applications from $\overline{2, j}$ groups interrupted by applications from groups $\overline{1, j-1}$: $\frac{K_{m-1}}{K_r - R_{m-1}} \rho(m,n)$ - probability of staying in queue (m, n) - applications, interrupted by applications from $\overline{1, m-1}$ groups. This probability is determined by the formula (8), taking into account the intensity $\lambda(m,n)$ of the flow (m, n) -payments;

$\eta_0(j,k) = \sum_{m=1}^{j} \sum_{n=1}^{N_m} \frac{K_r \rho_0}{K_r - R_{m-1}} \rho(m,n)\Delta(m,n)$ – average time of subscription of applications from $\overline{1, j}$ groups interrupted by device refusals;

$\frac{K_r \rho_0}{K_r - R_{m-1}} \rho(m,n)$ – the probability that the queue has (m, n)-applications, interrupted by the denial of the device. This probability is determined on the basis of (9) with account $\lambda(m,n)$;

$z_H(j,k)$ - average waiting time (j, k) – application, equal to the sum of the considered components without accounting σ_0;

$R_{j,k-1} = \sum_{m=1}^{j-1} \sum_{n=1}^{N_m} \rho(m,n) + \sum_{n=1}^{k-1} \rho(j,n)$.

Note that in each queue there can be no more than one application interrupted by applications with absolute priority or denial.

After simple transformations from equation (12) we obtain the following recurrence relation:

$$w_H(j,k) = \frac{1}{K_r - R_{j,k}} \left[K_r^2 \rho_0 \Delta_0 + \sum_{m=1}^{j} \sum_{n=1}^{N_m} \frac{1}{K_r - R_{m-1}} \times \right.$$

$$\times \rho(m,n)\Delta(m,n) + \sum_{m=1}^{j-1} \sum_{n=1}^{N_m} w_H(m,n)\rho(m,n) + \quad . \quad (13)$$

$$\left. + \sum_{n=1}^{k-1} w_H(j,n)\rho(j,n) \right]$$

Where $R_{j,k} = \sum_{m=1}^{j-1} \sum_{n=1}^{N_m} \rho(m,n) + \sum_{n=1}^{k} \rho(j,n)$.

To obtain a formula for explicit determination, we analyze the relation (13) for "pure" service disciplines with a relative and absolute priority.

For the discipline of service with a relative priority we receive:

- for the first flow $w_H(1,1) = \dfrac{K_r^3 \rho_0 \Delta_0 + \sum_{n=1}^{N_1} \rho(1,n)\Delta(1,n)}{K_r[K_r - \rho(1,1)]}$,

- for the second flow.

$$w_H(1,2) = \frac{K_r^3 \rho_0 \Delta_0 + \sum_{n=1}^{N_1} \rho(1,n)\Delta(1,n)}{[K_r - \rho(1,1)] \times [K_r - \rho(1,1) - \rho(1,2)]}.$$

These formulas allow us to assume a general solution in the form:

$$w_H(1,k) = \frac{K_r^3 \rho_0 \Delta_0 + \sum_{n=1}^{N_1} \rho(1,n)\Delta(1,n)}{(K_r - R_{1,k-1})(K_r - R_{1,k})}, \qquad (14)$$

Where $R_{1,k-1} = \sum_{n=1}^{k-1} \rho(1,n), \qquad R_{1,k} = \sum_{n=1}^{k} \rho(1,n)$.

For the discipline of service with absolute priority ($M = N, \quad N_m = 1$ for all $m = \overline{1,M}$) of the expression (13) we obtain:

• for the flow of the first group

$$w_H(1,1) = \frac{K_r^3 \rho_0 \Delta_0 + \rho(1,1)\Delta(1,1)}{K_r[K_r - \rho(1,1)]};$$

• for the flow of the second group

$$w_H(2,1) = \frac{K_r^3\rho_0\Delta_0 + \rho(1,1)\Delta(1,1) + \rho(2,1)\Delta(2,1)}{[K_r - \rho(1,1)][K_r - \rho(1,1) - \rho(2,1)]}.$$

Then on the basis of these equalities we get the general expression:

$$w_H(j,1) = \frac{K_r^3\rho_0\Delta_0 + \sum_{m=1}^{j}\rho(m,1)\Delta(m,1)}{(K_r - R_{j-1,1})(K_r - R_{j,1})}. \quad (15)$$

Where

$$R_{j-1,1} = \sum_{m=1}^{j-}\rho(m,1), \qquad R_{j,1} = \sum_{m=1}^{j}\rho(m,1).$$

Analyzing the expression (14) and (15), it is easy to assume the general form of the formula for determining $w_H(j,k)$ for a mixed discipline of service:

$$w_H(j,k) = \frac{K_r^3\rho_0\Delta_0 + \sum_{m=1}^{j}\sum_{n=1}^{N_m}\rho(m,n)\Delta(m,n)}{(K_r - R_{j,k-1})(K_r - R_{j,k})}. \quad (16)$$

Substituting formula (16) in (13) and making simple transformations, we can verify the validity of this assumption.

By expressions (11) and (16) we calculate the required average time of stay (*j*, *k*) -request *v* (*j*, *k*) in the AS-I system

Similarly, as for the system type AC-I, formulas can be derived for determining the temporal characteristics for the remaining systems type AC-II, BD-I, BD-II.

The models take into account the physical properties of the CC, such as instantaneous elasticity (dynamic allocation and release of resources for fast scaling according to needs) and measuring service (management and optimization of resources with the help of measuring instruments).

References

1. Matov A.Ya. Optimization of the provision of computing resources with adaptive cloud infrastructure**.** Data recording, storage and processing. 2018. T.20, No. 3 P.83-90. Ukr.
2. Matov A.Ya., Shpilev V.N., Komov A.D. et al. Organization of computational processes in ACS. Ed. A.Ya.Matov. Kiev, 1989. - 200s.Russ.
3. Matov, A.Ya., Khramova, I.O. Problems of mathematics and mathematical modeling of old ones are counted for the integration of information and analysis of the system and power management. Data recording, storage and processing. 2010. V. 12, №2. Pp. 113-127. Ukr.

4. Matov A. Ya. Two modes of continuous completion of a queue when the instrument is restored in a servicing system with a relative priority. Avtomat. i Telemekh., 66-70, 1974.
5. Matov A. Ya. Two priority system with an unreliable device and period of servicing. Engineering Cybernetics 10 (5), 849-852, 1973.
6. Matov A. Ya. Two continuous queue disciplines for service-resumption period in a nonpreemptive-priority queuing system. Automation and remote control 35 (4), 575-578, 1974
7. Matov A.Ya., Tishchenko N.F. Mathematical models of computing systems with priority denial of service. Izv. Academy of Sciences of the USSR. Technical cybernetics. - 1980. - №3. - p. 190-194. Russ.
8. Matov A.Ya., Shpilev V.N. The use of combined priorities to improve the efficiency of computing processes in the ACS. Mechanization and automation of management. - 1983. - №4. - p. 58-60. Russ.
9. Matov A Ya, Zhluktenko VI, Chernous KA, Tishchenko NF. Two continuous queuing disciplines in mixed priority systems. Cybernetics and Systems Analysis 14 (3), 1978. C. 421-426.
10. Mokrov E.V., Samuilov K.E. Cloud computing system model in the form of a queuing system with multiple queues and with a group of requests. https://cyberleninka.ru/article/n/model-sistemy-oblachnyh-vychisleniy-v-vide-sistemy-massovogo-obsluzhivaniya-s-neskolkimi-ocheredyami-i-s-gruppovym-postupleniem-zayavok. Russ.
11. Tsai J.M., Hung S.W. A novel model of technology diffusion:system dynamics perspective for cloud computing // Journal of Engineering and Technology Management. 2014. V. 33. P. 4762. doi: 10.1016/j.jengtecman.2014.02.003
12. Singh P., Dutta M., Aggarwal N. A review of task scheduling based on meta-heuristics approach in cloud computing // Knowledge and Information Systems. 2017. V. 52. N 1. doi: 10.1007/s10115-017-1044-2

FUZZY SET OBJECTS CLUSTERING METHOD USING EVOLUTION TECHNOLOGIES

Hnatiienko Hryhorii Mykolaiovych[1] and Suprun Oleh Oleksiiovych[2]
Intellectual and Information Systems Department, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine,
[1]g.gna5@ukr.net, [2] oleh.o.suprun@gmail.com

Presented in the article method allows an expert, or person, who makes decisions, to effectively evaluate, calculate and make the best decision on the large set of opinions. The biggest problem while performing the trustful quality assessment is to build an adequate comparison matrix, based on human opinions, since reviews may vary greatly, and it's impossible to appreciate them all, when dealing with big data. Thus, a method to rank and cluster these results is required. Analyzing the fuzzy set objects, most of classic algorithms can not be used, since it's impossible or hard to design a proper target function, so the evolution technology method is used. It allows not only to build a proper clusters around the given centers, but also to find these centers, to evaluate the best radius and the most important attributes. Besides that, large amount of parameters allows an expert to choose the most important aspects or vary the time, used to solve the problem. The algorithm and required preparations are described in the article.

Keywords: *Complex Information Systems, Fuzzy Sets, Evolution Technologies, Objects Clustering*.

Introduction

In the modern information society, more and more information require evaluating, investigating and ranking every day. It can be seen in every day life, when planning the day or the nearest future, choosing different products in the shop, or looking for the best options of summer vocation. In all these situations the human has to overview large amounts of information and responds to make right decision, in most cases it is done subconsciously.

The same problem, but in much greater scale, appears while planning the future steps of big companies, or even countries, choosing the best products to improve and simulating the economical and financial models. The consequences are very important, affect the future of many people. But, to make right decisions, it's necessary to take into account the thoughts of many experts, their opinions and advises, the number of such experts in big companies may vary from dozens to hundreds.

Besides the great number of opinions, very often they differ not only by their content, but also the form, and a lot of time is required to build a normalized table of opinions. To handle this automatically the fuzzy set principles are used. But, at the same time, most algorithms of data mining are made to evaluate strict statistical data, that has the same form and ranged values, that can't be done with expert opinions. So, the non-classic method is required to simulate the work of a team that ranks and makes basic conclusions, based on large number of opinions.

1 Clustering the Fuzzy Set Objects

Cluster analysis is a method of grouping experimental data into classes. Performing this method, the following condition must be fulfilled: the experimental values or data that are gathered into one class or group are closer to each other than to values from other classes, according to a certain parameter or attribute. The number of clusters can be arbitrary or constant. The main purpose of data clustering methods is to ensure that the similarity of the data that is combined into one cluster is maximal. As a result of solving the clusterization problem, the values ranking in relatively similar groups is performed. The cluster analysis application is very diverse and common in many subject areas.

Clustering methods can be used to construct a fuzzy set membership function. By definition, membership function $\mu_A(x)$ quantitatively calibrates the membership of the fundamental set elements of the considerations space $x \in X$ to the fuzzy set $\tilde{A}$. The value $\mu_A(x) = 0$ means that the fuzzy set does not include an element x. The value $\mu_A(x) = 1$ means the full membership of an element to a fuzzy set. The values of the membership function from the interval $(0,1)$ numerically characterize fuzzy elements.

The designing of membership functions (belonging function, F-functions) is one of the most important stages in solving decision-making problems in the fuzzy statement. The uncertainty of measuring the attribute intensity of an object can consist of the complexity of measurement, inaccurate intensity measurement, different perception of the objects properties by experts, etc. The membership function must fulfil the following requirements:

- continuity, that is a formalization of the following intuitive statement: when two solutions of the set X do not differ much from each other, the values of membership functions for these solutions are also closeto each other;

- consistency with the ratio of advantages, i.e. $\mu_D(x_1) \geq \mu_D(x_2)$ then and only when $x_1 \succ x_2$.

Let a series of empirical data in a range $(0,1)$, obtained as a result of measuring a certain value x is given. According to the data analysis results, it is necessary to formulate a conclusion acceptable to the researcher, what values the indicated variable acquires. One-dimensional analysis suggests describing the distribution of one variable, including its central trend (including average values, median and mode), and dispersion (including the range and quintile of the data set, and dissemination measures such as variance and standard deviation). Since for many practical tasks the determination of the mode, the median, or the average mean of a given series is insufficient, a cluster analysis can be applied to the specified series. It is necessary to define clusters that can be used to build an membership function for the value x, which best classify the results of measuring the value x and allow to design an membership function for the value x in the interval $(0,1)$.

The membership function of a fuzzy set is a generalization of the indicator function of the classical set. In fuzzy logic, it is a degree of truthfulness. The degrees of truth are sometimes confused with probability, but they are fundamentally different, since truth indicates that the value belongs to a given set that is not similar to any phenomenon or condition. Algorithms for determining the membership function by analyzing the frequency of values were considered in [1, 2].

Fuzziness appears when the expert tries to quantify the subject area. Obtaining fuzzy knowledge is an extremely difficult task, such as experts, as a rule, are not able to adequately design a membership function. Therefore, so-called standard membership functions, which have a given form and are described by well-known analytic functions, are often used. Standard membership function features are easy to apply to many practical tasks.

2 Clusterization Algorithm

When solving multicriteria optimization problems, the problem of determining the Pareto domain is strictly objective and solved without the use of any heuristics. Narrowing the area of effective objects requires the use of additional information from experts, since effective set of parameters can not be compared formally with each other. As a rule, three heuristics are used to determine a single solution to a multicriteria problem:

- one of the allowed transformations is used to transform all values of object parameters to dimensionless form in a given value range;
- the vector of the criteria relative importance is determined;
- it is assumed that the multicriteria problem solution is the point of intersection of the normalized weight coefficients of criteria relative importance beam with the field of effective alternatives to the problem.

It is known that designing a structured table of benefits in a formalized form is a complex task for a human. Research in the field of expert evaluation tasks and the practice of building decision support systems show that far not always experts and decision makers have a clear idea of the structure of preferences for a plurality of objects. In most cases, a person can not adequately determine the weight factors, or allocate in the obvious case the heuristics that are used in a decision-making situation.

A common method of presenting the weight coefficients values for n objects with index i, $i \in I = \{1,...,n\}$ are valid numbers taking into account the condition of normalization:

$$\sum_{i \in I} \rho_i = 1,\ \rho_i > 0,\ i \in I.$$

Let as the result of a calculation series the weight coefficients set is formed:

$$\rho_i \in \left\{ \rho_i^1, ..., \rho_i^L \right\}, i \in I,$$

where L is the number of values indexes of normalized weighting coefficients obtained as a computations result. Based on the obtained values, the membership functions of the weight coefficients values in the fuzzy set (0,1) are determined. Approaches to the determination of membership functions and algorithms for constructing membership functions based on the analysis of the values frequency are given in [2].

3 Evolution Technologies Implementing

Using the elements of evolutionary technologies for clustering fuzzy sets objects is logical and can give practically useful results, since they allow to avoid most problems, that the classic methods meat. For example, in most cases the target function is undifferentiated, it may not have a certain interpretation, since it's built upon the human opinions, that are rather hard to normalize.

The most common situation while evaluating the expert's conclusions is the need to analyse their thoughts on the set of objects according to more then one attribute. Thus, the clusterization problem can be stated as follows: to split the set Ω, that consists of n objects into m clusters $\Omega = \{S_1, S_2, ..., S_m\}$. Every object S_i has the set of characteristics, i, $i \in I = \{1, ..., n\}$. The objects data are presented in the "object-attribute" type table:

TABLE I. OBJECTS DATA

X / S	X_1	X_2	...	X_k
S_1	ρ_{11}	ρ_{12}	...	ρ_{1k}
S_2	ρ_{21}	ρ_{22}	...	ρ_{2k}
...	...	...	...	...
S_n	ρ_{n1}	ρ_{n2}	...	ρ_{nk}

Let the set $C^* = \{C_1^*, C_2^*, ..., C_m^*\}$ be the solution of the clustering problem. Considering that the objects are points located inside a k-dimensional hyperparallelepiped, the following can be obtained:

$$C^* = \arg\min_{C \in \theta} F(c) = \arg\min_{C \in \theta} \min_{(C_1, C_2, ... C_m)} \sum_{i=1}^{n} \sum_{j=1}^{m} \chi\{S_i \in R_j\} \cdot d(S_i, C_j),$$

with restriction that all potential cluster centers lie inside the hyperparallelepiped, $C = (C_1, C_2, ..., C_m)$.

To find the problem solution, genetic algorithm can be used, since it is one of the most flexible evolution method. The solution search can be presented as a sequence of the following steps:

Step 1. Generate q sets, that consist of m elements $C^i = (C_1^i, C_2^i, ..., C_m^i), i = \overline{1,q}$ and, as usual, q is a number from the set $\{20, 21, ..., 50\}$. The values C_j^i are uniformly distributed in the hypercube, $i = \overline{1,q}$, $j = \overline{1,m}$,

Step 2. Calculate the distance from each object to each cluster center:

$$d_{jp}^{i} = d\left(S_j, C_p^j\right) = \left(\sum_{l=1}^{k}\left(x_{jl} - c_{pl}^{j}\right)^2\right)^{1/2},$$

where $i = \overline{1,q}$, $j = \overline{1,n}$, $p = \overline{1,m}$.

In order to determine which cluster the objects belong to, the search problem must be solved: for $\forall i = \overline{1,q}$, $\forall j = \overline{1,n}$ find

$$\arg\min_{p} d(S_j, C_p^i).$$

Table 2 is obtained, where p_{ij} is a cluster number, which the object S_j belongs to, for i-th potential problem solution.

TABLE II. OBJECT-CLASTER ATTACHMENT

i / S	S_1	S_2	…	S_k
1	p_{11}	p_{12}	…	p_{1n}
2	p_{21}	p_{22}	…	p_{2n}
…	…	…	…	…
q	p_{q1}	p_{q1}	…	p_{qn}

Step 3. Calculate the distance form every object to the center of appropriate cluster that is the potential problem solution:

$$d_i = \sum_{j=1}^{n} d(S_j, C_{p_{ij}}^{i}), \forall i = \overline{1,q}.$$

Step 4. If the Genetic Algorithm is chosen to solve the problem, then, considering the value d_i, the following operations are made with the set C^i: crossover operation, mutation, and elite selection into the new population of potential solutions is made.

If the Evolution Strategy method is used, the new potential population is generated, where every new solution is obtained from "parent" solution, by adding a normally distributed random displacement

$X_{new} = X_{parent} + \xi(N(0,\delta^2))$. The amount of new potential solutions, as usual, is 7 times bigger, than the amount of "parent" solutions [1, 2]. The best solutions from the "parent" and intermediate populations are selected to the new population.

Step 5. Steps 3-5 steps 3-5 are repeated until the criteria for iterative process stop are not achieved. Such criteria may be:

– the priori number of iterations;

– for a given ε :

$$\left|\max_i d_i^{(it)} - \max_i d_i^{(it+1)}\right| < \varepsilon ,$$

or

$$\left|avg\ d_i^{(it)} - avgd_i^{(it+1)}\right| < \varepsilon ,$$

or

$$\max_i d(C^{i(it)}, C^{i(it+1)}) < \varepsilon ,$$

where it is iteration number.

Step 6. The clustering problem solutions, after the criteria for iterative process stop are achieved, are the following:

$$\arg\min_{C^{i(it)}} d_i^{(it)} .$$

The proposed clustering method is a parametric method and it's efficiency depends on researcher's qualification and efficiency of parameters setting for each specific problem. These parameters are the following:

- The iteration process stop criteria;
- The crossover and mutation type;
- The type of parents selection and population of the next generation formation, if the Genetic Algorithm is chosen as optimization method;
- The parent-solutions and offspring-solution number;
- Constant or variable value standard deviation;
- Positive or negative dynamics of standard deviation for the Evolution Strategy.

It is known that the convergence in probability holds for a genetic algorithm with an elite selection to a new solutions population, and for

$(\lambda + \mu)$ – Evolution Strategy, where λ - parent-solution number, and μ – offspring-solution number with iterations number tending to infinity.

Conclusions

Working with big data is one of many aspects, connected with modern information society, and nowadays a lot of ranking and clustering methods are designed and improved. They are required to evaluate and compute large amount of statistical data, like economical or financial reports and other. But the main flaw of most of them is that these methods can be used only with the same systematic data.

At the same time, for large companies it is necessary to combine, or at least rank the reviews and opinions of many experts, that is impossible using the classic methods.

To solve this problem the fuzzy sets are used, They allow to formalize and combine different opinions and to make some conclusions, based on them. The main principles of performing the fuzzy set objects, so that the proper results can be obtained, and the main heuristics, are described in the article. This allows to normalize the results, so that more complicated algorithm can be implemented.

Clustering different objects can be mat in different fields of science, but it's rather new for the fuzzy sets. The target function may differ very much, it can be undifferentiated and discontinuous, and thus the classic methods can't be used. The evolution algorithm performance is presented, that allows to conduct the clustering process in permissible time period.

As for the future investigations, the possibility of genetic algorithm improvement must be considered. For example, the usage of evolution strategy, or implementing the penalty method to improve the clusterization results.

References

1. Hnatyenko H.M. Algorithms for determining the membership function by analyzing the values frequency // Proceedings of the III-th international school-seminar "Theory of decision-making", Uzhorod, 2006. – pp. 32-34.
2. Hnatyenko H.M., Snityuk V.E. Expert Decision Technology: Monograph. - K .: LLC "Maclaut", 2008. – 444 p.

СЕМАНТИЧЕСКИЕ МОДЕЛИ В ЗАДАЧЕ МОНИТОРИНГА ОБЩЕСТВЕННОГО МНЕНИЯ

Додонов А.Г., Ландэ Д.В., Березин Б.А.
Институт проблем регистрации информации НАН Украины, Киев, Украина

Предложен метод построения и использования семантических моделей (СМ) с целью непрерывного во времени мониторинга общественного мнения (МОМ). Семантические модели позволяют использовать при мониторинге общественного мнения результаты лингвостатистического анализа текстов (Text Mining), применения методов извлечения информации (Information Extraction), оценки тональности (sentiment Analysis). Предложенная процедура МОМ включает три основные этапа: построение и кластеризацию СМ; отбор документов и определение их тональности; визуализацию результатов. Предложено построение СМ с помощью алгоритма компактифицированного графа горизонтальной видимости (CHVG), применение методов кластерного анализа для определения актуальных тематик, оценивание доли и тональности отдельных кластеров в составе общего тематического потока информации. Рассмотрены модели отдельных предметных областей. Результаты анализа подтверждают возможность использования предложенного метода мониторинга общественного мнения в различных предметных областях.

Ключевые слова: *модель предметной области, семантическая модель, кластерный анализ, анализ тональности, контент-мониторинг, мониторинг общественного мнения*

1 Постановка проблемы

Под семантической моделью (СМ) в рамках данной работы будем понимать модель предметной области, имеющую вид ориентированного графа, вершины которого соответствуют концептам предметной области, а дуги задают отношения между ними. Такая семантическая модель может трактоваться как семантическая карта предметной области. Семантические модели позволяют использовать при мониторинге общественного мнения

результаты лингвостатистического анализа текстов (Text Mining), применения методов извлечения информации (Information Extraction), содержащейся в текстах из сети Интернет. В то время, как существующие проекты анализа общественного мнения больше ориентированы на разовые (статичные) исследования общественного мнения относительно объектов и явлений, в данной работе предлагается метод автоматизированного построения и использования СМ на основе непрерывного во времени мониторинга общественного мнения в сети Интернет. Предлагается проведение анализа общественного мнения на основе методов обработки естественного языка (Natural Language Processing). Такой анализ направлен на определение отношения субъекта мониторинга общественного мнения к выбранной теме. Одной из основных задач анализа общественного мнения является классификация эмоциональной окраски текста (положительной, отрицательной или нейтральной).

В обзорных работах, посвященных анализу, извлечению мнений, настроений (Sentiment Analysis – SA, Opinion Mining – OM) отмечается, что это компьютерное изучение мнений, отношения людей к объекту, концепту, которая может представлять личности, события или темы [1,2]. В этих работах выделяются уровни анализа мнений: уровень документа, уровень предложения и аспектный уровень, когда рассматривается мнение в отношении некоторого концепта. В [3] анализируется общественное мнение о выборах президента США в 2012 г. на основе новостных статей, опубликованных в Интернет. На основе этих ресурсов из сети Интернет были выделены триплеты “субъект-глагол-объект” и с помощью их построены семантические графы. Результаты анализа избирательной кампании были получены путем исследования характеристик семантических графов.

2 Метод построения и использования семантических моделей

В данной работе предложен метод построения и использования СМ для задач МОМ в сети Интернет, предусматривающий три этапа:

- построение и кластеризация СМ;
- отбор документов и определение тональности тематик;
- визуализация результатов.

На первом этапе производится:

- выборка массива документов для построения СМ;
- нахождение концептов; определение связей СМ путем

построения компактифицированного графа горизонтальной видимости [4];
- кластеризация графа;
- формирование запросов, соответствующих кластерам (на основе найденных кластеров экспертами выделяются тематики и формулируются запросы для отбора соответствующих документов).

На втором этапе производится:
- отбор документов, соответствующих тематикам (подтемы), из общего информационного потока с помощью запросов;
- определяется их доля в общем потоке документов;
- определяется тональность документов соответствующих тематик.

На третьем этапе тематики с тональностями:
- визуализируются на карте;
- записываются состояния в базу данных (БД) системы мониторинга для последующего получения динамики изменения результатов во времени.

Ниже рассмотрены основные операции, выполняемые в составе этих трех этапов.

3 Этап построения и кластеризации СМ

Этап предусматривает отбор массива документов для построения семантической модели. На основе заданного объекта мониторинга формулируется запрос на выборку массива документов.

Нахождение концептов. Входящие в массив документы проходят предварительную обработку, удаление служебной информации, а также стоп-слов, не несущих смысловой нагрузки. Также проводится стемминг (приведение слов к основе). Затем, на основе учета известных метрик (например, $TFIDF$) из слов массива документов отбираются наиболее важные, имеющие наибольший вес понятия.

Определение связей СМ путем построения графа горизонтальной видимости. Для определения связей между концептами и построения семантической модели используется алгоритм компактифицированного графа горизонтальной видимости (КГГВ) [7].

Особенность использования алгоритма КГГВ в данной работе состоит в том, что его первые два шага выполняются отдельно для каждого предложения анализируемого текста. OBORB процессе

разработки предложенного метода, проводилось исследование построения СМ для документов, собираемым по темам Шелкового пути («One Belt, One Road» – OBOR), Норд Стрим-2, ГМО и некоторым другим. Фрагмент графа семантической модели, построенной для 28 концептов темы OBOR с помощью описанного алгоритма приведен на Рис. 1.

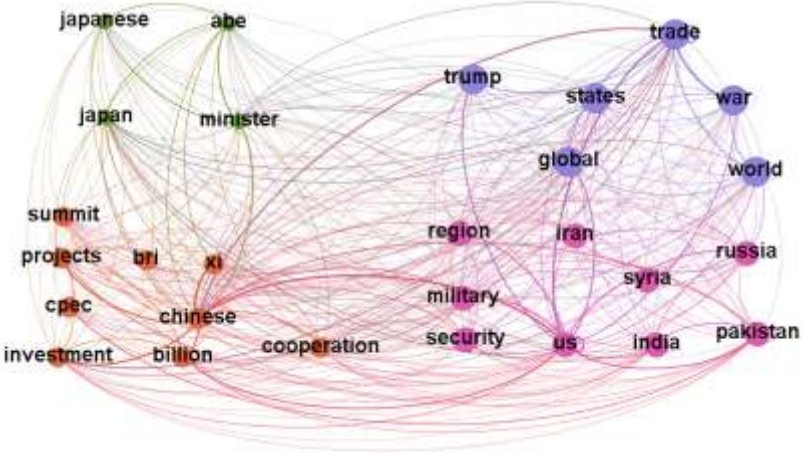


Рис. 1. Фрагмент графа семантической модели для 28 концептов темы OBOR

Кластеризация графа СМ. Учитывая актуальность аспектного уровня анализа мнений, после построения семантической модели анализируется ее сетевая структура с помощью алгоритмов кластеризации графов – выявления сообществ (clastering graph, community detection). Сообщество определяется как плотно связанная группа узлов, которая слабо связана с остальной частью сети. В [5] рассмотрено более десятка алгоритмов кластеризации для выявления сообществ как непересекающихся, так и перекрывающихся (и сообществ обоих типов). Для кластеризации графов СМ рассматривалось применение различных известных алгоритмов. Среди известных алгоритмов поиска сообществ в графе можно выделить алгоритм Louvain, в соответствии с которым в начале алгоритма каждая вершина образует отдельное сообщество.

Формирование запросов, соответствующих кластерам. В результате кластеризации графа семантической модели были найдены множества наиболее связанных вершин графа, соответствующих выявленным кластерам, т.е. множества близких

понятий. На основе этих понятий в общем потоке документов, характеризующих анализируемую предметную область, выделяются тематики, аспекты. Эксперты в данной предметной области дают названия этим тематикам и формулируют запросы для последующего отбора с помощью информационно-поисковой системы документов, соответствующих тематикам, из общего потока документов, характеризующих общую предметную область.

Для рассматриваемой темы OBOR**,** на основе найденных кластеров, экспертами были сформулированы четыре тематики и соответствующие им запросы, приведенные в Табл. 1. На этом, периодически повторяемый этап обучения системы мониторинга (период повторения от нескольких часов до суток), реализуемый при помощи алгоритма построения и кластеризации СМ, заканчивается.

Этап отбора документов и определения тональности тематик. Отбор документов тематик из общего информационного потока с помощью запросов. Из общего потока документов, формируемого по поисковому запросу, характеризующему предметную область, отбираются документы тематик с помощью поисковых запросов, сформулированных экспертами на основе выявленных в семантической модели кластеров. Для отобранных документов каждой из тематик определяется их доля в общем потоке документов.

Для потока документов, формируемого по поисковому запросу к системе контент-мониторинга InfoStream ***(one-road)&(one-belt)&china***, характеризующего тему OBOR**,** названия тематик и соответствующие запросы приведены в Таблице 1. Доли документов тематик, отобранных с помощью сформулированных запросов из общего потока, приведены в таблице 3.

Определение тональности документов тематик. Для документов каждой из выявленных тематик определяется тональность – позитивная, негативная, нейтральная на основе анализа слов, входящих в состав документов, относящихся к темам. Под тональностью текста в данном случае понимается позитивная, негативная или нейтральная эмоциональная окраска как всего текстового документа, так и отдельных его частей, имеющих отношения к определенным понятиям, таким как персоны, организации, бренды и т. п. В задаче определения тональности проверяется как минимум три показателя эмоциональной окраски: позитивная, негативная, нейтральная и, зачастую, существует потребность также в проверке комбинации этих гипотез (например, для выявления уровня «экспрессивности» текста). Тональности

документов тематик, отобранных из общего потока документов темы OBOR приведены в Таблице 2.

Таблица 1. Тематики и запросы, сформулированные на основе найденных кластеров

Номер	Ключевые слова тематики, запрос	Название тематики
1	Xi projects investment	Председатель КНР (президент) Си Цзиньпин об инвестиционных проектах в составе инициативы bri
2	India Pakistan US	Отношение Индии, Пакистана, США и др стран к инициативе
3	south region development	Отношение стран к развитию южного региона в рамках инициативы
4	Japan minister Abe	Премьер министр Японии Абэ об инициативе

Таблица 2. Доля документов сформулированных тематик в общем потоке темы «OBOR» и их тональность

Тематика	Доля	Негативная	Нейтральная	Позитивная
Xi projects investment	10.1% (101)	2% (2)	0	98% (99)
Japan minister Abe	9.2% (92)	1% (1)	0	99% (91)
India Pakistan US	8% (80)	5% (4)	0	95% (76)
south region development	15.3% (153)	5% (8)	1% (1)	94% (144)

Этап визуализации результатов. На данном этапе выполняется визуализация найденных тематик с тональностями на карте. Результаты мониторинга визуализируются в режиме реального времени на географической карте с привязкой к конкретным объектам. По каждой выявленной в общем потоке документов тематике отображается диаграмма с указанием названия тематики и доли документов общего потока, приходящейся на эту тематику, а

также с отображением доли документов положительной, отрицательной и нейтральной тональности внутри тематик. На географической карте показываются: тематики, выявленные в потоке входных документов; доля документов по каждой тематике; тональность документов по тематикам, а также динамика изменения результатов во времени. Найденные в процессе мониторинга состояния записываются в БД мониторинга для последующего получения динамики изменения результатов во времени.

Операции, выполняемые в составе трех этапов рассмотренного метода реализованы с помощью средств программного пакета Gephi (http://gephi.org), а также с помощью программных средств, разработанных на языке программирования для статистических расчетов R. Результаты использования предложенного метода приведены ниже.

4 Полученные результаты

Возможности применения предложенного метода построения и использования СМ для мониторинга общественного мнения анализировались на основе результатов мониторинга Интернет-ресурсов по нескольким темам:

- One Belt, One Road (OBOR) - инициатива Китайской Народной Республики по объединённым проектам «Экономического пояса Шёлкового пути» и «Морского Шёлкового пути XXI века».
- Nord Stream – проект газопровода из России в Германию через Балтийское море.
- GMO – генетически модифицированные организмы и некоторые другие темы. Для мониторинга общественного мнения по теме OBOR анализировался массив из 1000 англоязычных документов (период с 30.11.2018 по 25.07.2018) собранных при помощи запроса ***(one-road)&(one-belt)&china*** с использованием системы InfoStream.

На первом этапе, после выборки массива документов и его предварительной обработки, нахождения концептов (на основе частоты использования терминов, а также на основе показателя *TFIDF*) были построены соответствующие СМ (Рис. 1).

На втором этапе, доли документов тематик, отобранных с помощью сформулированных запросов из общего потока, приведены в Таблице 2. Там же приведены тональности документов

сформулированных тематик, отобранных из общего потока документов темы OBOR**.**

На третьем этапе выполняется визуализация полученных результатов. Динамика изменения доли документов и тональности сформулированных тематик в составе документов темы OBOR по неделям показана на Рис. 2.

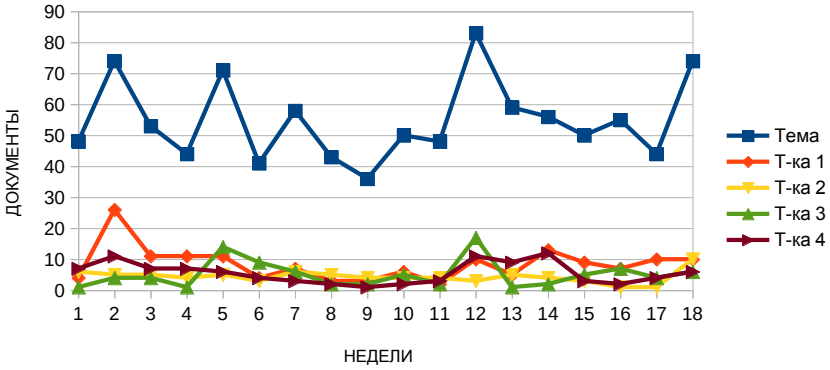


Рис. 2. Изменение доли документов сформулированных тематик (Тематика 1 — Тематика 4, четыре нижних графика) в составе документов темы OBOR (Тема, верхний график) по неделям

Кроме темы OBOR**,** рассматривалось использование предложенного метода мониторинга общественного мнения для тем Nord Stream, GMO и других. Например, для темы Nord Stream, на основе массива из 1000 англоязычных документов, (собранных в период с 02.11.2018 по 18.08.2018) была построена соответствующая СМ, найдены кластеры и сформулированы тематики: merkel putin meeting (о встрече Меркель и Путина); gas transit ukraine (о транспортировке газа через Украину); european security energy market (о безопасности европейского энергетического рынка), poland united states (об отношении Польши и США к проекту Nord Stream).

Выводы

Предложен метод построения и использования СМ для мониторинга общественного мнения, включающий три этапа: построение и кластеризацию СМ; отбор документов и определение тональности тематик; визуализацию результатов.

Показано построение СМ с помощью алгоритма компактифицированного графа горизонтальной видимости, применение методов кластерного анализа для определения актуальных тематик, оценивание доли и тональности отдельных подтем в составе общего тематического потока информации.

Полученные результаты подтверждают возможность использования предложенного метода мониторинга общественного мнения в различных предметных областях.

Литература

1. *Schouten K., Frasincar F.* Survey on aspect-level sentiment analysis // IEEE Transactions on Knowledge and Data Engineering, 2016. – Iss. 28(3). – pp. 813-830.
2. *Medhat W., Hassan A., Korashy H.* Sentiment analysis algorithms and applications: A survey // Ain Shams Engineering Journal, 2014. – Iss. 5(4). – pp. 1093-1113.
3. *Sudhahar S., Veltri G., Cristianini N.* Automated analysis of the US presidential elections using Big Data and network analysis // Big Data & Society, 2015. – Iss. 2(1). – pp. P. 21-49.
4. *Lande D.V., Snarskii A.A., Yagunova E.V., Pronoza E.V.* The use of horizontal visibility graphs to identify the words that define the informational structure of a text // 12th Mexican International Conference on Artificial Intelligence (MICAI), 2013. – pp. 209-215. DOI: 10.1109/MICAI.2013.33
5. *Harenberg S., Bello G., Gjeltema L., Ranshous S., Harlalka J., Seay R., Samatova N.* Community detection in large-scale networks: a survey and empirical evaluation // Wiley Interdisciplinary Reviews: Computational Statistics, 2014. – Iss. 6(6). – pp. 426-439.

ОБҐРУНТУВАННЯ ВИБОРУ НЕПЕРЕРВНИХ КАРТ ДЛЯ ПРЕДСТАВЛЕННЯ ОЦІНОК РИЗИКІВ ІНФОРМАЦІЙНОЇ БЕЗПЕКИ

Володимир Мохор[1], Олександр Бакалинський[2], Василь Цуркан[3]

[1]Інститут проблем моделювання в енергетиці ім. Г.Є. Пухова Національної академії наук України, Київ, Україна
[2] Департамент формування та реалізації державної політики у сфері кіберзахисту Державної служби спеціального зв'язку та захисту інформації України, Київ, Україна
[3] Інститут спеціального зв'язку та захисту інформації Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна
v.mokhor@gmail.com, baov@meta.ua, v.v.tsurkan@gmail.com

Розглянуто представлення оцінок ризиків інформаційної безпеки картою. Приділено увагу особливостям такого представлення. Встановлено обмеження дискретності та нерівномірності кроку зміни значень величини ризику. Для подолання встановлених обмежень запропоновано використання неперервних карт ризиків. Це обумовлено надходженням подій інформаційної безпеки неперервним потоком з огляду на аналогію системи управління інформаційною безпекою з системою масового обслуговування. Тому *обґрунтування неперервних карт зведено до оцінювання імовірності появи подій з прийнятним ризиком. Для цього використано поняття та методи геометричної імовірності. Завдяки цьому отримано "одиничний квадрат" як відображення геометричного місця точок, що відповідають значенням нормованої величини ризику інформаційної безпеки. Межу прийнятності представлено гіперболою. З огляду на це обґрунтовано використання неперервних карт ризиків.*

Ключові слова: *інформаційна безпека, ризик, оцінювання ризику, неперервна карта ризиків.*

Вступ

Системне, наочне представлення оцінок ризиків інформаційної безпеки здійснюється шляхом використання карт. Традиційно вони відображаються координатною площиною, осями якої є параметри

ризику. Здебільшого, такими параметрами є імовірність реалізації загрози та величина втрат [1] - [4].

Однак, на практиці використання карт обмежується їх дискретністю і нерівномірністю кроку зміни значень величини ризику [2], [3].

Основна частина

Для подолання обмежень дискретності і нерівномірності зміни значень величини ризику пропонується використання неперервних карт [2]. Вони отримуються унаслідок збільшення кратності дискретних карт ризиків. Такий перехід обумовлений надходженням подій інформаційної безпеки неперервним потоком з огляду на аналогію системи управління інформаційною безпекою з системою масового обслуговування. Тому цей процес найкраще відображається неперервними картами ризиків. Тоді як їх результативність підтверджується прикладами застосування у медицині [2], [4].

Обґрунтування неперервних карт зводиться до оцінювання імовірності появи подій з прийнятним ризиком. Для цього використовуються поняття та методи геометричної імовірності. Вводиться двомірна декартова система координат по горизонтальній осі якої відкладено значення імовірності реалізації загрози, вертикальній – значення величини втрат. Узгодженість зміни значення обох діапазонів досягається нормуванням величини втрат. Завдяки цьому в декартовій системі координат отримується “одиничний квадрат”. Цією фігурою відображається геометричне місце точок, що відповідають вірогідним значенням нормованої величини ризику інформаційної безпеки. Шляхом задання прийнятного її значення здійснюється поділ множини ризиків на підмножини прийнятних і неприйнятних. Межа прийнятності представляється функціональною залежністю у формі гіперболи. З огляду на це, обґрунтовується використання неперервних карт для представлення оцінок ризиків інформаційної безпеки.

Висновок

Обґрунтування використовуваності неперервних карт оцінок ризиків інформаційної здійснено на основі, по-перше, аналогій системи управління інформаційною безпекою з системою масового обслуговування; по-друге, поняття і методів геометричної імовірності. Це дозволяє подолати встановлені обмеження

дискретності та нерівномірності зміни значень величини ризику для карт.

Список використаної літератури

1. International Organization for Standardization. (2011, June 10). ISO/IEC 27005. Information technology. Security techniques. Information security risk management, https://www.iso.org/standard/56742.html.
2. Мохор, В., Бакалинський, О., Цуркан В.: Представлення оцінок ризиків інформаційної безпеки картою ризиків. Information Technology and Security, vol. 6, iss. 2, 94–104 (2018), doi: 10.20535/2411-1031.2018.6.2.153494.
3. Астахов, А.: Исскуство управления информационными рисками. ДМК Пресс. Москва (2010).
4. Mokhor, V., Bogdanov, A., Bakalinskii, A., Tsurkan, V.: The Method of the Design Requirements Formation for Information Security Management System, Selected Papers of the XVI International Scientific and Practical Conference “Information Technologies and Security”. Kyiv, 2016, pp. 1-6, http://ceur-ws.org/Vol-1813/paper7.pdf.

РІЗНОВИДИ МАНІПУЛЯТИВНИХ ФОРМ ВИКОРИСТАННЯ СОЦІАЛЬНОЇ ІНЖЕНЕРІЇ У КІБЕРПРОСТОРІ

Оксана Цуркан, Ростислав Герасимов

[1] Інститут проблем моделювання в енергетиці ім. Г.Є. Пухова Національної академії наук України, Київ-164, Україна
o.tsurkan24@gmail.com, gerasimov.rostislav@gmail.com

Розглянуто особливості використання соціальної інженерії в кіберпросторі. Встановлено маніпулятивність її впливу на свідомість (підсвідомість) людини. Для цього застосовуються властиві їй слабкості, потреби, манії (пристрасті), захоплення. Маніпулювання ними дозволяє отримати "санкціонований" доступ до інформації без руйнування та перекручування головних для людини системоутворюючих якостей. При цьому виділено такі різновиди маніпулятивних форм як шахрайство, обман, афера, інтрига, містифікація, провокація. Обман як маніпулятивна форма дозволяє отримати "санкціонований" доступу до конфіденційної інформації шляхом умисного введення в оману людини. Основними засобами обману розглядаються брехня, напівправда, замовчування. Крім цього, розглянуто аферу та провокацію як маніпулятивні форми отримання "санкціонованого" доступу до конфіденційної інформації за заздалегідь визначеним сценарієм та проведення спонукаючи спланованих дій, у другому. Наведено приклади використання маніпулятивних форм у кіберпросторі.

Ключові слова: *кіберпростір, соціальна інженерія, маніпулятивна форма, фішинг, фармінг, смішінг, вішінг, прітекстінг.*

Вступ

Використання соціальної інженерії в кіберпросторі полягає в маніпулятивному впливові на свідомість (підсвідомість) людини через властиві їй вразливості [1]. Зокрема, її слабкості, потреби, манії (пристрасті), захоплення. Маніпулювання ними дозволяє отримати несанкціонований доступ до інформації без руйнування та перекручування головних для людини системоутворюючих якостей [2].

Маніпулювання вразливостями людини виражається в таких формах як, наприклад [3], шахрайство, обман, афера, інтрига, містифікація, провокація.

Основна частина

Обман – отримання “санкціонованого” доступу до конфіденційної інформації шляхом умисного введення в оману людини. Основними засобами обману є брехня, напівправда, замовчування. Наприклад [4], [5], масове розсилання електронної пошти великій групі адресатів (фішінгу). Ознайомлення з електронними листами спонукає їх до, наприклад, відкриття вкладення до листа, переходу за посиланням на веб-сторінку; перенаправлення користувачів на шахрайські сайти для отримання їх логіну та паролю (фармінгу). Це досягається завдяки розповсюдженню електронної пошти серед користувачів, наприклад, соціальних мереж, онлайн-банкінгу, поштових веб-сервісів; отримання інформації шляхом масового розсилання SMS повідомлень з посиланням на веб-ресурси або з реквізитами організацій (смішінг). Внаслідок цього здійснюються відповідні дії, наприклад, дзвінок до банку для перевірки стану рахунку з зазначенням конфіденційних даних: номеру картки, терміну дії; отримання інформації шляхом входження в довіру під час розмови через ір-телефон (вішінг).

Афера – отримання “санкціонованого” доступу до конфіденційної інформації за заздалегідь визначеним сценарієм. Провокація – проведення спонукаючих спланованих дій, що призводять до отримання “санкціонованого” доступу до конфіденційної інформації. Наприклад [4], [5], отримання інформації або спонукання до вчинення певних дій обманом на основі заздалегідь складеного сценарію або створення фіктивної ситуації (прітекстінг). Застосовується через телефон та потребує проведення попередніх досліджень для входження в довіру.

Висновок

Різновиди маніпулятивних форм призводять до впливу на свідомість або підсвідомість людини проти волі, але за її згодою. В кінцевому випадку це спонукає до нової моделі поведінки людини та створення сприятливих обставин використання соціальної інженерії в кіберпросторі.

Так, обман взято за основу фішингу, фармінгу, смішінгу, вішінгу. Тоді як створення фіктивних ситуацій при прітекстінгу здійснюється завдяки афері та провокації.

Список використаної літератури

1. Мохор, В., Богданов, А., Килевой, А.: Наставления по кибербезопасности (ISO/IEC 27032:2013). ООО «Три-К», Киев (2013).
2. Winterfeld, S., Andress, J.: The Basics of Cyber Warfare: Understanding the Fundamentals of Cyber Warfare in Theory and Practice. Elsevier, Waltham (2013).
3. Tsurkan, O., Herasymov, R.: Detection of vulnerabilities of the computer systems and networks using social engineering techniques. Information Technology and Security, vol. 6, iss. 2, 43–50 (2018), doi: 10.20535/2411-1031.2018.6.2.153494.
4. Mokhor, V., Tsurkan, O., Tsurkan, V., Herasymov, R.: Information Security Assessment of Computer Systems by Socio-engineering Approach, Selected Papers of the XVII International Scientific and Practical Conference "Information Technologies and Security". Kyiv, 2017, pp. 1-6, http://ceur-ws.org/Vol-2067/paper13.pdf.
5. Mouton, F., Leenen, L., Venter, H.: Social engineering attack examples, templates and scenarios, Computers & Security, vol. 59, 186–209 (2016), doi: 10.1016/j.cose.2016.03.004.

ТЕОРЕТИКО-ГРАФОВИЙ ПІДХІД ДО МОДЕЛЮВАННЯ АТАК КОМП'ЮТЕРНИХ МЕРЕЖ

Володимир Булдижов

Компанія "СМК", Київ-2, Україна
vladimir.univer@gmail.com

Розглянуто підходи до моделювання атак комп'ютерних мереж. Серед них виокремлено теоретико-графовий підхід. Його використання зводиться до моделювання способів поєднання уразливостей для виконання атаки комп'ютерної мережі. При цьому використовуються набори, пов'язаних зі захистом перед- та післяумов. Завдяки цьому отримується граф атак, яким відображаються вірогідні сценарії атак зловмисника або шкідливого програмного забезпечення. На основі вірогідних сценаріїв атак визначаються найбільш небезпечні вразливості, атаки та засоби усунення уразливостей. Граф атак відображено направленим графом з множинами уразливостей і станів комп'ютерної мережі з урахуванням перед- і післяумов. З огляду на це визначено два різновиди відношень: кон'юнктивне (для використання уразливості необхідні всі стани як передумови), диз'юнктивне (для отримання визначеного стану післяумови достатньо використання будь-якої уразливості). У кінцевому випадку формуються настанови запобігання вірогідним атакам комп'ютерних мереж.

Ключові слова: *комп'ютерна мережа, дані, уразливість, атака, захист даних, моделювання атак, граф атак, теоретико-графовий підхід.*

Вступ

Одним з ключових компонентів інфраструктури інформаційних технологій є комп'ютерні мережі. Тому захист даних мереж від атак має велике значення як для окремих організацій, так і держави в цілому [1]. Порушення захищеності комп'ютерних мереж викликається різноманітними причинами, наприклад [2] - [5], наявністю вразливостей у програмних застосунках; помилок налаштування апаратного та програмного забезпечення; шкідливих програмних застосунків.

Одним із найбільш розповсюджених підходів до моделювання уразливостей комп'ютерних мереж є теоретико-графовий [1] - [5].

Основна частина

Побудова графів атак дозволяє моделювати способи поєднання уразливостей для виконання атаки комп'ютерної мережі. При цьому використовуються набори, пов'язаних зі захистом, умов, наприклад [1], наявність уразливості у визначеному вузлі або між вузлами комп'ютерної мережі. З огляду на це, використання теоретико-графового підходу дозволяє відобразити вірогідні сценарії атак зловмисника або шкідливого програмного забезпечення. Шляхом аналізування отриманих сценаріїв визначаються, наприклад [1]:

- перелік найбільш небезпечних уразливостей;
- перелік атак;
- перелік засобів усунення уразливостей;

Граф атак представляється направленим графом G ($V \cup S$, $require \cup provides$), де $V \cup S$ – множина вершин; $require \cup provides$ – множина дуг; V – множина уразливостей; S – множина станів комп'ютерної мережі (перед- і післяумов); $require \subseteq V \times S$ – кон'юнктивне відношення (для використання уразливості необхідні всі стани як передумови); $require \subseteq S \times V$ – диз'юнктивне відношення (для отримання визначеного стану післяумови достатньо використання будь-якої уразливості).

Завдяки моделюванню атак комп'ютерних мереж за теоретико-графовим підходом формуються настанови запобігання вірогідним атакам [1], [5]. Ці настанови орієнтовані як на зміну архітектури комп'ютерної мережі, так і на встановлення оновлень програмного забезпечення.

Висновок

Моделювання атак комп'ютерних мереж за теоретико-графовим підходом дозволяє, по-перше, визначити їх вірогідні сценарії; по-друге, обрати засоби усунення вразливостей; по-третє, зіставити обрані засоби з вірогідними сценаріями атак і, як наслідок, по-четверте, сформувати настанови запобігання атакам комп'ютерних мереж.

Список використаної літератури

1. Булдижов, В.: Электротехнические аналоги в оценке рисков информационной безопасности, Электронное моделирование, т. 34, № 6, 1–6 (2012).

2. Котенко, И.В., Левшун, Д.С., Чечулин, А.А., Ушаков, И.А., Красов А.В.: Комплексный подход к обеспечению безопасности киберфизических систем на основе микроконтроллеров, Вопросы кибербезопасности, № 3 (27), 29–38 (2018).
3. Дойникова, Е.В., Котенко, И.В.: Методики и программный компонент оценки рисков на основе графов атак для систем управления информацией и событиями безопасности, Информационно-управляющие системы, № 5, 54–65 (2016).
4. Котенко, И.В., Степашкин, М.В.: Анализ защищенности компьютерных сетей на основе моделирования действий злоумышленников и построения графа атак, Труды ИСА РАН, том 31, 126–207 (2007).
5. Колегов, Д.Н.: Проблемы синтеза и анализа графов атак, SecurityLab.ru (2007), https://www.securitylab.ru/contest/299868.php.

Содержание

Национальная академия наук Украины
Институт проблем регистрации информации

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И БЕЗОПАСНОСТЬ

Материалы XVIII Международной научно-практической конференции

Выпуск 18

Підп. до друку 18.12.2018. Ум. друк. арк. 14,65. Обл.-вид. арк. 25,66.
Наклад 100 пр. Зам. № 15-200.

ТОВ "Інжиніринг"
ISBN 978-966-2344-69-1