doi: https://doi.org/10.15407/visn2017.03.045

ЛАНДЕ
Дмитро Володимирович —
доктор технічних наук,
завідувач відділу спеціалізованих засобів моделювання Інституту проблем реєстрації інформації НАН України

АНАЛІЗ ІНФОРМАЦІЙНИХ ПОТОКІВ У ГЛОБАЛЬНИХ КОМП'ЮТЕРНИХ МЕРЕЖАХ

За матеріалами наукової доповіді на засіданні Президії НАН України 25 січня 2017 року

У доповіді наведено результати досліджень з розроблення фундаментальних і прикладних основ аналізу інформаційних потоків у глобальних комп'ютерних мережах. Обґрунтовано актуальність цього завдання, показано параметри сучасного інформаційного простору, існуючі теоретичні і технологічні рішення. Наведено опис методологічних та інструментальних засобів аналізу інформаційних потоків, розроблених в Інституті проблем реєстрації інформації НАН України, зокрема моделювання інформаційних потоків, розподіленого контент-моніторингу глобальних мереж, формування багатомовних повнотекстових баз даних, аналізу динаміки тематичних інформаційних потоків із застосуванням вейвлет- і фрактального аналізу, автоматичного формування моделей предметних областей.

***Ключові слова*:** інформаційні потоки, контент-моніторинг, вейвлет-аналіз, фрактальний аналіз, моделі предметних областей.

Вступ

У глобальних мережах міститься величезна кількість інформації, за допомогою якої можна вирішувати найрізноманітніші завдання. Однак при цьому важливо вибрати саме ту інформацію, яка цікавить користувача.

Велика кількість інформаційних ресурсів у глобальних мережах містить різні експертні оцінки, певна частина яких пов'язана з реалізацією інформаційних впливів, здійсненням спрямованих інформаційних операцій, веденням інформаційних війн. Такі матеріали можуть бути проаналізовані, узагальнені, на їх основі можна створювати бази для подальшого прийняття рішень, які відрізняються від традиційних експертних оцінок як за обсягами, так і за рівнем об'єктивності.

Крім того, в мережах може розміщуватися інформація, пов'язана з організацією протиправної діяльності, тероризмом. Відомі також ефекти так званої *мережевої мобілізації*, впливу

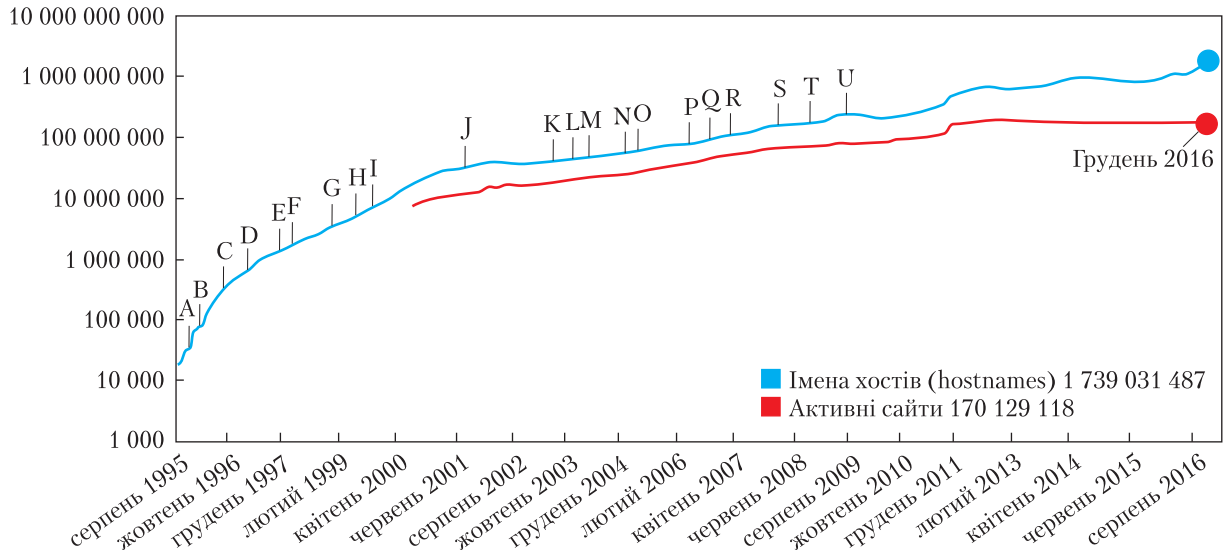

Рис. 1. Динаміка кількості сайтів в усіх доменах (за даними Netcraft, логарифмічна шкала)

на людську свідомість, управління і маніпулювання громадською думкою. Аналізуючи інформаційні потоки, можна знайти і так звані *інформаційні резервації* — частини інформаційного простору, які характеризуються замкненістю, обмеженістю тематики даних.

Отже, врахування інформації з мережевих джерел відіграє важливу роль як для виявлення напрямів розвитку економіки, науки, технологій та інших сфер життя, так і для вирішення конкретних завдань у сферах безпеки людини, суспільства, держави.

Деякі параметри інформаційного простору

На сьогодні кількість інформаційних ресурсів у глобальних мережах перевищує сотні трильйонів документів. У 2014 р. система Google вже індексувала в мережі 60 трлн документів, причому ці документи розміщуються не лише на веб-сайтах. За даними відомого інтернет-сервісу Netcraft (http://netcraft.com), у 2014 р. кількість веб-сайтів у мережі вже перетнула позначку в мільярд, а нині становить понад 1,7 млрд (рис. 1).

У 2014 р. у виданні Supercomputing Frontiers and Innovations з'явилась публікація, в якій стверджувалося, що приблизний обсяг даних у мережі Інтернет сягає 10^{24} байтів, тобто один йотабайт. Лише в одній соціальній мережі Facebook активні користувачі генерують більш як 4 петабайти даних за добу. За даними компанії CISCO, обсяг інтернет-трафіку в 2016 р. досяг одного зеттабайта, тобто 1 099 511 627 776 гігабайтів.

Технологічні рішення

При збиранні й аналізі таких даних виникають проблеми, пов'язані з *обробкою надвеликих обсягів даних, пошуком і навігацією в динамічних інформаційних потоках*. Величезна кількість багатомовних інформаційних ресурсів зумовлює складність їх використання при здійсненні інформаційно-аналітичної роботи. Для вирішення цих проблем сьогодні застосовуються такі технологічні концепції, як Big Data (великі дані, рис. 2), Complex Networks (складні мережі), Cloud Computing (хмарні обчислення), Data/Text Mining (глибинний аналіз даних і тексту).

Проблеми розмірності і динаміки багатомовних інформаційних ресурсів у глобальних мережах потребують проведення фундаментальних досліджень у галузі дискретної мате-

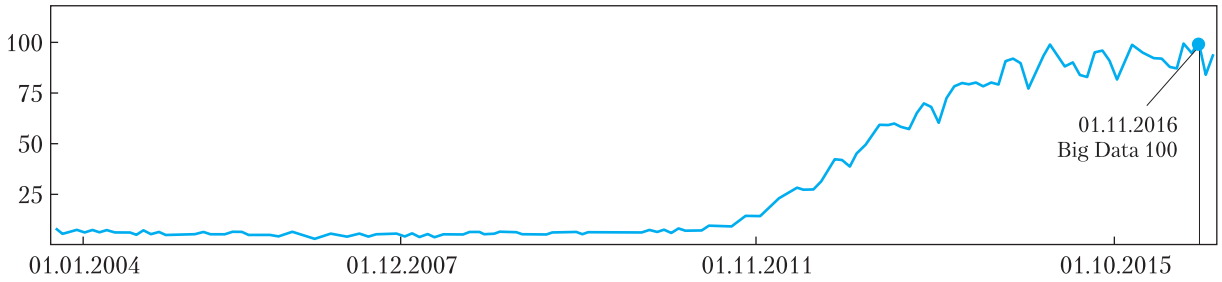

Рис. 2. Аналіз терміна Big Data за допомогою Google Trends

матики (теорії графів, мереж), розпізнавання образів (класифікація, кластерний аналіз), лінгвістики, цифрової обробки сигналів, вейвлет- і фрактального аналізу тощо.

У світі й досі залишаються невирішеними завдання ефективної аналітичної обробки інформації з глобальних мереж, оперативного вилучення необхідних фактографічних даних, виявлення трендів в окремих предметних областях, розпізнавання змістових аномалій, прогнозування тощо. Більшість із зазначених завдань — це актуальні проблеми семантичної обробки надвеликих динамічних масивів інформації. Навіть спроби часткового практичного вирішення цих проблем зумовили успішність таких проектів, як пошукові системи Google, Yandex, Baidu, системи моніторингу соціальних мереж (SMM) типу Keyhole, Brandwatch, CyberAlert, аналітичні системи типу Palantir, Centrifuge, i2 та ін. В Україні до таких систем можна віднести, зокрема, системи, створені на базі наукових результатів, отриманих в Інституті проблем реєстрації інформації (ІПРІ) НАН України, — система контент-моніторингу InfoStream, аналітична система X-SCIF, система сканування ресурсів соціальних медіа Robusta тощо.

Розробки Інституту проблем реєстрації інформації НАН України

В ІПРІ НАН України було теоретично обґрунтовано і створено засоби:

• моделювання інформаційних потоків у глобальних комп'ютерних мережах, зокрема мультиагентну модель розповсюдження інформації;

• розподіленого контент-моніторингу глобальних мереж;

• формування багатомовних повнотекстових баз даних;

• аналізу динаміки тематичних інформаційних потоків, зокрема, вперше застосовано вейвлет-аналіз до задач виявлення інформаційних операцій;

• прогнозування розвитку подій на основі фрактального аналізу;

• формування мереж взаємозв'язку понять, що екстрагуються із тестових масивів, і аналізу цих мереж;

• автоматичного формування моделей предметних областей.

Мультиагентна модель розповсюдження інформації. Для моделювання тематичних інформаційних потоків як полігон для подальших досліджень в ІПРІ НАН України створено мультиагентну модель поширення інформації в соціальних мережах [1, 2]. Для цього формується близький до реальності віртуальний інформаційний простір, населений віртуальними агентами, з якими асоціюються окремі повідомлення в соціальній мережі і які інкапсулюють у собі гіперпосилання на інформаційні ресурси мережі Інтернет. Передбачається, що окремі агенти можуть самозароджуватися; породжувати нових агентів шляхом репостингу (repost); «вмирати» — зникати з простору агентів; отримувати лайки (like) від інших агентів. Кожен агент має «потенціал», залежний від часу його життя, авторитетності (гіперпосилань, проставлених на нього) і пло-

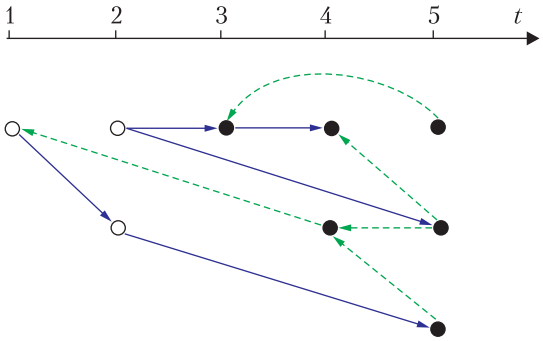

Рис. 3. Приклад динаміки мультиагентної системи

дючості (кількості породжених безпосередньо ним агентів).

Варіювання відповідними параметрами моделі дає можливість змоделювати різноманітні профілі поведінки інформаційних сюжетів. На рис. 3 наведено приклад можливої динаміки мультиагентної системи.

У результаті проведених досліджень було реалізовано програму еволюції простору агентів, досліджено еволюцію мультиагентної системи, знайдено аналогії з реальними тематичними інформаційними потоками. Виявлено статистичні закономірності, що стосуються життєвого циклу окремих повідомлень, розподіл яких відповідає розподілу Вейбулла. Дані моделювання було підтверджено шляхом порівняння з реальною мережею мікроблогів Twitter.

Мережа інформаційних проксі-серверів. Найпоширеніша причина відмов від надання веб-сайтами свого контенту за запитами користувачів полягає в їх перевантаженості. При цьому є досить жорсткі обмеження можливостей веб-сайтів у разі масової роботи з їх контентом. Слід зауважити, що більшість цих обмежень не враховується навіть у нормативних документах, що регламентують вимоги стосовно захисту інформації на веб-сторінках. Такі ситуації, як зловмисна DoS-атака, кризова пікова відвідуваність призводять до недоступності інформаційних ресурсів веб-сайтів, зокрема, для аналітиків та осіб, які приймають рішення. При цьому системи моніторингу та аналізу інформаційних потоків у мережі Інтернет не повинні бути джерелом перевантаження цільових веб-ресурсів.

Як підхід до вирішення зазначених проблем в ІПРІ НАН України пропонується побудова мережі — системи пов'язаних між собою інформаційних проксі-серверів [3]. При цьому до даних, які обслуговує інформаційний проксі-сервер, висуваються такі вимоги:

• враховується динамічна новинна складова веб-простору як найкритичніша з точки зору забезпечення оперативного доступу;

• множина веб-сайтів для кешування вибирається експертами відповідно до внеску цих джерел в інформаційний простір;

• інформація в проксі-сервері представляється в універсальному внутрішньосистемному форматі, який передбачає однозначне синтаксичне трактування (наприклад, XML);

• дані в інформаційному сховищі мають оновлюватися за розкладом, що відповідає динаміці їх поновлення на веб-сайтах.

Проксі-сервер, з одного боку, призначений для надійного обслуговування користувачів корпоративних мереж, а з іншого — може забезпечувати обмін даними з аналогічними зовнішніми проксі-серверами.

Система багатомовних повнотекстових баз даних. Система багатомовних повнотекстових баз даних, яка створюється на базі концепції Big Data з використанням універсальної системи кодування (UTF), включає такі технологічні засоби [4]:

1) збирання та первинної обробки інформації, формування багатомовних вхідних інформаційних потоків, наведених у національних кодуваннях (рис. 4), для чого має забезпечуватися формування макроописів інформаційних ресурсів, що відповідають різним мовам;

2) виявлення опорних слів за статистичними алгоритмами в інформаційних матеріалах, наведених різними мовами;

3) тематичної рубрикації документів;

4) створення, ротації баз даних і забезпечення формування внутрішніх словникових наборів даних;

5) створення та підтримки баз даних, що відповідають різним мовам, у тому числі формування окремих пошукових індексів баз даних, які охоплюють слова, наведені різними мовами;

Рис. 4. Інтерфейс адміністратора багатомовних інформаційних ресурсів

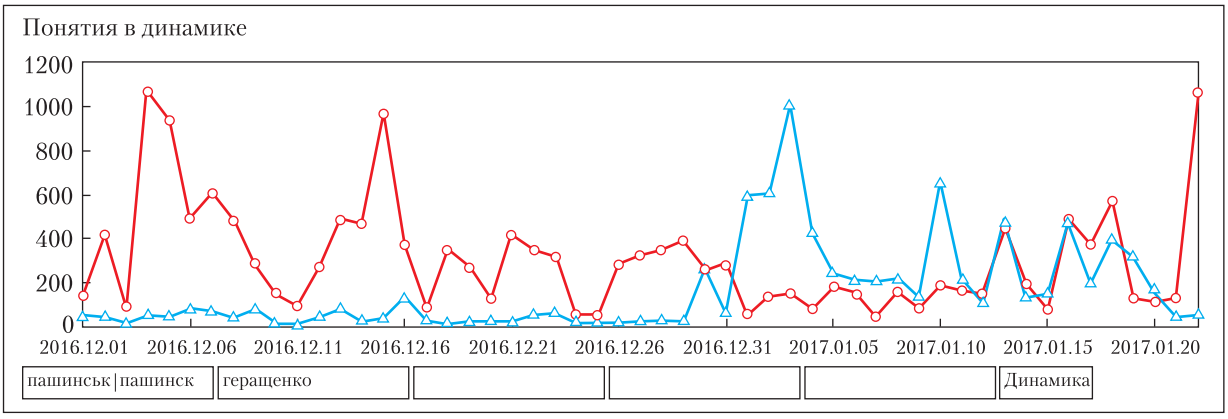

Рис. 5. Відображення динаміки інформаційних потоків у системі InfoStream

6) пошуку в індексах баз даних, що охоплюють слова, наведені різними мовами;

7) формування аналітичних звітів, у тому числі інформаційних портретів і сюжетних ланцюжків, що ґрунтуються на використанні опорних слів, наведених різними мовами.

Аналіз динаміки тематичних інформаційних потоків. Тематичним інформаційним потокам можна ставити у відповідність часові ряди (інтенсивність публікацій в одиницю часу, рис. 5) [5], для аналізу яких все частіше обґрунтовано застосовуються формальні методи: статистичного, фрактального, фур'є- або вейвлет-аналізу.

Аналіз цих числових рядів надає можливості виявлення трендів, циклів, аномалій, подальшого прогнозування розвитку інформаційних сюжетів, пов'язаних з вибраними тематиками, порівняння числових рядів, що відповідають різним інформаційним сюжетам.

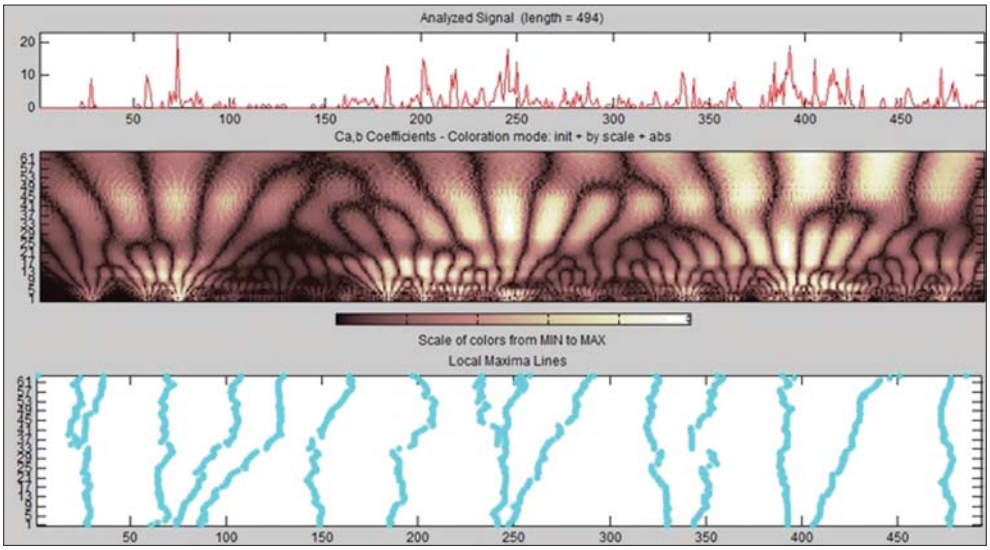

Рис. 6. Динаміка публікацій за цільовою тематикою і вейвлет-скейлограма (вейвлет Морле) інформаційного потоку

Застосування вейвлет-аналізу. В ІПРІ НАН України вперше запропоновано застосування вейвлет-аналізу для дослідження тематичних інформаційних потоків [6, 7].

Технологія використання вейвлетів дозволяє виявляти одиничні та нерегулярні «сплески», різкі зміни значень кількісних показників у різні періоди часу, зокрема обсягів тематичних публікацій в Інтернеті. При цьому можуть виявитися моменти виникнення циклів, а також моменти, коли за періодами регулярної динаміки настають хаотичні коливання.

Періодичні зміни, які відбуваються для значень коефіцієнтів вейвлет-перетворення, на деякій безперервній множині частот мають вигляд ланцюжка «пагорбів» з вершинами в точках (по осі часу), в яких ці зміни досягають найбільших значень. Іншим важливим показником є тенденція динаміки часового ряду (тренд) незалежно від періодичних коливань. Наявність тренду може бути неочевидною при простому розгляді часового ряду, наприклад, якщо тренд поєднується з періодичними коливаннями. Тренд відображується на скейлограмі як плавна зміна яскравості вздовж осі часу одночасно на всіх масштабах. Якщо тренд зростаючий, то яскравість збільшується, якщо спадний — зменшується. Крім того, важливим фактором, який необхідно враховувати при аналізі часових рядів, є локальні особливості, тобто можливі різкі, стрибкоподібні зміни характеристик вихідного ряду.

На відповідній вейвлет-скейлограмі (рис. 6) можна бачити всі характерні особливості вихідного ряду: масштаб та інтенсивність періодичних змін, напрям і значення трендів, наявність, розташування, тривалість локальних особливостей.

Виявлення інформаційних операцій. Як методологічну основу виявлення інформаційних операцій запропоновано дослідження динаміки інформаційних потоків [7]. Здійснено системні дослідження такого багатоаспектного явища, як інформаційні операції. На рис. 7 наведено узагальнену діаграму, що відповідає етапам життєвого циклу інформаційних операцій.

Вейвлет-аналіз є одним з ефективних засобів виявлення в тематичних інформаційних потоках шаблонів, що відповідають наведеному на рис. 7, у різних масштабах. Визначено, що динаміку інформаційних операцій найточніше відображують такі відомі вейвлети, як «мексиканський капелюх» та вейвлет Морле.

Прогнозування розвитку подій на основі фрактального аналізу. Для дослідження часових рядів обсягів повідомлень у тематичних

інформаційних потоках сьогодні все ширше використовується теорія фракталів, традиційна сфера застосування якої — фрактальна геометрія, обробка зображень тощо. Проте часові ряди, породжувані тематичними інформаційними потоками, також мають фрактальні властивості і їх можна розглядати як стохастичні фрактали. Такий підхід розширює сферу застосування теорії фракталів на інформаційні потоки, динаміка яких описується засобами теорії випадкових процесів [8].

З іншого боку, теорія фракталів розглядається як підхід до статистичного дослідження, що дозволяє одержувати важливі характеристики інформаційних потоків, не вдаючись у детальний аналіз їхньої внутрішньої структури і зв'язків.

Найважливішою характеристикою рядів, що мають хаотичну поведінку, є, як відомо, фрактальна розмірність, яку в багатьох випадках можна обчислити за допомогою так званого *R/S*-аналізу. Точніше кажучи, обчислюється не сама фрактальна розмірність, а показник Херста (*H*), пов'язаний з нею простим співвідношенням. *R/S*-аналіз ґрунтується на аналізі розкиду значень досліджуваного ряду *R* до середньоквадратичного відхилення *S*. Числові значення *H* характеризують різні типи кореляційної динаміки (персистентності). При $H = 0{,}5$ спостерігається некорельована поведінка значень ряду, а значення $0{,}5 < H < 1$ відповідають рівню автокореляції ряду.

Автоматичне формування мереж взаємозв'язків понять і моделей предметних областей. За допомогою автоматичних методів екстрагування фактичних даних, що застосовуються в системах інтеграції інтернет-ресурсів, формуються так звані «інформаційні портрети», що охоплюють списки опорних слів, імен персон, топонімів, мов, компаній і т. ін., що містяться в документах, релевантних деякому тематичному запиту. Ці списки можуть агрегуватися, в результаті чого можливе визначення взаємозв'язків окремих понять [9].

Сьогодні під моделлю предметної області, зокрема, розуміють спеціальним чином сформовану мережу понять, онтологію.

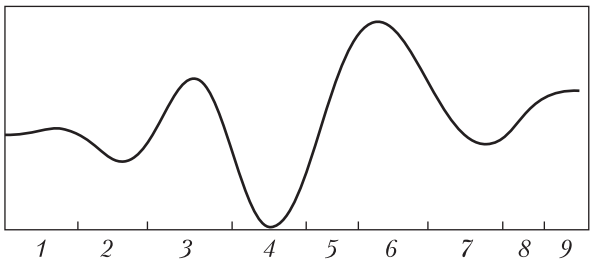

Рис. 7. Життєвий цикл інформаційних операцій: *1* — фон; *2* — тиша; *3* — «артпідготовка»; *4* — тиша; *5* — атака/тригер зростання; *6* — пік підвищених очікувань; *7* — втрата ілюзій організаторів; *8* — суспільне усвідомлення; *9* — продуктивність/фон

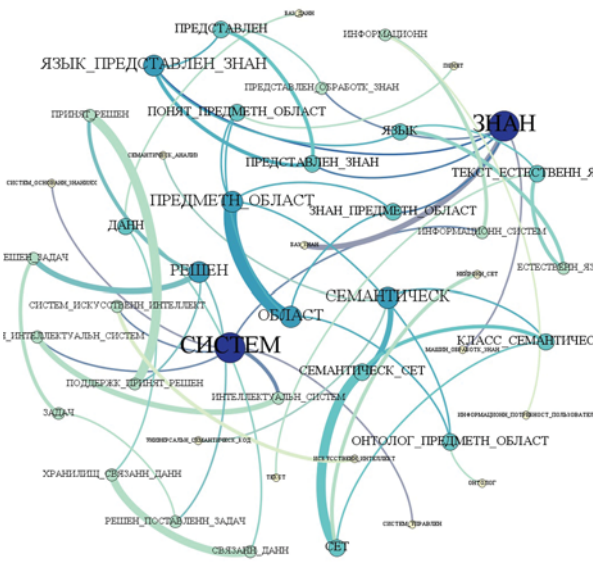

Рис. 8. Візуалізація фрагмента моделі предметної області

Як термінологічна основа для формування термінологічних онтологій використовуються «інформаційні портрети». Після автоматичного виявлення зв'язків між окремими поняттями формуються мережі, що відповідають моделям предметних областей [10], і здійснюється їх відображення програмними засобами аналізу і візуалізації графів (рис. 8).

Сценарний аналіз на базі застосування когнітивних карт. Сценарії інформаційної підтримки, протистояння деструктивним впливам пов'язані з певними поняттями предметної області. На базі *x* моделей предметної області за-

безпечується формування когнітивних карт [11, 12], вершинам яких відповідають деякі поняття (концепти), а ребрам (зв'язкам) — причинно-наслідкові (каузальні) зв'язки між концептами. Після вибору цільових понять у когнітивних картах виділяються підграфи, які найбільш тісно пов'язані з цільовими об'єктами. При формуванні когнітивних карт вирішується декілька змістових завдань, серед яких:

1) виявлення вузлів — концептів;

2) виявлення семантичних (каузальних) зв'язків між вузлами;

3) ранжування концептів, виявлення головних із них;

4) ранжування зв'язків;

5) виявлення концептів, що впливають на цільові вузли-об'єкти;

6) виявлення підграфів, найбільш тісно пов'язаних з цільовими об'єктами;

7) візуалізація і редагування когнітивних карт;

8) формування сценаріїв;

9) прогнозування розвитку ситуацій у результаті реалізації сценаріїв.

Впровадження

На основі наведених розробок створюються системні, програмні й технологічні рішення, які використовуються для визначення нових напрямів науки і техніки, розвитку конкурентоспроможності галузей економіки, при автоматизації заходів з інформаційної безпеки. На сьогодні є вже десятки впроваджень запропонованих рішень, зокрема, при створенні таких систем, як система автоматизованого збору та розподілу інформації з веб-сайтів мережі Інтернет у Службі безпеки України, система контент-моніторингу відкритих вебресурсів мережі Інтернет у внутрішній мережі для Служби зовнішньої розвідки України, Програмно-технічний комплекс інтегрованого доступу до новинної інформації в Управлінні справами Апарату Верховної Ради України. У 2016 р. створено спільну лабораторію — ІПРІ НАН України і ІПСА НТУУ «КПІ імені Ігоря Сікорського» (українська сторона) і Академії наук провінції Шаньдун (китайська сторона) — для виконання робіт за відповідною тематикою.

За результатами досліджень з цього напряму, проведених в Інституті проблем реєстрації інформації НАН України, опубліковано 17 монографій та понад 100 статей. Для апробації, координації та поширення теоретичних результатів у галузі комп'ютерного моделювання, моніторингу і аналізу інформаційних ресурсів, інформаційних потоків у глобальних комп'ютерних мережах, соціально-правового моделювання та інших інформаційних технологій в ІПРІ НАН України вже протягом 16 років проводяться щорічні міжнародні науково-практичні конференції «Інформаційні технології та безпека».

Отже, роботи з теорії і практики аналізу інформаційних потоків у глобальних комп'ютерних мережах, їх моніторингу і комп'ютерного моделювання є актуальними, необхідними для визначення науково-технічних напрямів, оцінки економічної, суспільно-політичної ситуації як в країні, так і в усьому світі, підготовки важливих управлінських рішень, зокрема в галузях інформаційної і кібернетичної безпеки.

REFERENCES

1. Lande D.V., Hraivoronska A.M., Berezin B.O. Model of information spread in social networks. *European Journal of Natural History.* 2016. (5): 41.
2. Lande D., Dodonov V., Kovalenko T. Corporate system of monitoring of network information resources on the basis of a multiagent approach. *Information Technology and Security*. 2016. **4**(1): 4.
[Ланде Д., Додонов В., Коваленко Т. Корпоративная система мониторинга сетевых информационных ресурсов на основе мультиагентного подхода. *Information Technology and Security*. 2016. Т. 4, № 1. С. 4—12.]
3. Dodonov A.G., Lande D.V. Organization of a network of information proxy servers. *Data Recording, Storage & Processing.* 2006. **8**(3): 24.

[Додонов А.Г., Ланде Д.В. Организация сети информационных прокси-серверов. *Реєстрація, зберігання і обробка даних*. 2006. Т. 8, № 3. С. 24—32.]
4. Dodonov A.G., Lande D.V., Putyatin V.G. *Computer networks and analytical researches*. (Kyiv, 2014).
[Додонов А.Г., Ландэ Д.В., Путятин В.Г. *Компьютерные сети и аналитические исследования*. К.: ИПРИ НАН Украины, 2014.]
5. Dodonov A.G., Lande D.V., Putyatin V.G. *Information flows on wide computer networks*. (Kyiv: Naukova Dumka, 2009).
[Додонов О.Г., Ланде Д.В., Путятін В.Г. *Інформаційні потоки в глобальних комп'ютерних мережах*. К.: Наук. думка, 2009.]
6. Lande D.V., Dodonov V.A. Non-linear properties of multiagent model of distribution of news. *Information Technology and Security*. 2016. **4**(2): 80.
[Ланде Д., Додонов В. Нелинейные свойства мультиагентной модели распространения новостей. *Information Technology and Security*. 2016. Т. 4, № 2. С. 80—92.]
7. Gorbulin V.P., Dodonov A.G., Lande D.V. *Information operations and security of society: threats, opposition, modeling*. (Kyiv: Intertechnology, 2009).
[Горбулін В.П., Додонов О.Г., Ланде Д.В. *Інформаційні операції та безпека суспільства: загрози, протидія, моделювання*. К.: Інтертехнологія, 2009.]
8. Lande D.V., Snarskii A.A., Bezsudnov I.V. *Internetics: Navigation on complex networks: models and algorithms*. (Moscow: Librocom (Editorial URSS), 2009).
[Ландэ Д.В., Снарский А.А., Безсуднов И.В. *Интернетика: Навигация в сложных сетях: модели и алгоритмы*. М.: Либроком (Editorial URSS), 2009.]
9. Dodonov A.G., Lande D.V. Detection of concepts and their correlations within technology of content monitoring. *Data Recording, Storage & Processing*. 2006. **8**(4): 45.
[Додонов А.Г., Ландэ Д.В. Выявление понятий и их взаимосвязей в рамках технологии контент-мониторинга. *Реєстрація, зберігання і обробка даних*. 2006. Т. 8, № 4. С. 45—52.]
10. Dodonov A.G., Lande D.V., Putyatin V.G. Domain model in organizational management systems. In: *Decision Support Systems. Theory and practice*: Proc. XI Conf. (8 June, 2015, Kyiv, Ukraine). P. 29—32.
[Додонов А.Г., Ландэ Д.В., Путятин В.Г. Модели предметных областей в системах организационного управления. В кн.: *Системы поддержки принятия решений. Теория и практика.(СППР'2015)*: матер. 11-й науч.-практ. конф. (8 июня 2015 г., Киев, Украина). С. 29—32.]
11. Snarskii A.A., Zorinets D.I., Lande D.V. "Conjectural" links in complex networks. *Physica A*. 2016. **462**: 266.
12. Dodonov A.G., Lande D.V., Boychenko A.V. Scenario approach in case of a research of dynamics of information flows on the Internet. In: *OSTIS-2015*: Proc. Int. Sci. Conf. (19—21 Feb., 2015, Minsk). P. 225—230.
[Додонов А.Г., Ландэ Д.В., Бойченко А.В. Сценарный подход при исследовании динамики информационных потоков в сети Интернет. В кн.: *Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2015)*: матер. V междунар. науч.-техн. конф. (19—21 февраля 2015, Минск). С. 225—230.]

D.V. Lande

Institute for Information Recording of National Academy of Sciences of Ukraine (Kyiv)

ANALYSIS OF INFORMATION FLOWS IN GLOBAL COMPUTER NETWORKS

According to the materials of scientific report at the meeting of the Presidium of NAS of Ukraine, January 25, 2017

The report presents the research results of the Institute for Information Recording on the development of basic and applied aspects of the analysis of information flows in global computer networks. The relevance of this task is justified; parameters of the modern information space, the existing theoretical and technology solutions are shown. The description of the methodology and tools for analysis of information flows, modeling of information flows, distributed content monitoring in global networks, formation of multilingual full-text databases, analysis of dynamics of subject information flows using wavelet and fractal analysis, automatic generation of domain models are provided.

Keywords: information flows, content monitoring, wavelet analysis, fractal analysis, subject areas models.