УДК 004.7

ПОБУДОВА МЕРЕЖІ ТЕРМІВ У СФЕРІ КІБЕРБЕЗПЕКИ ЗА ДАНИМИ СЕРВІСУ GOOGLE SCHOLAR

Д. В. Ланде[1, а], О. О. Дмитренко[1, б]

[1] *Інститут проблем реєстрації інформації НАН України*

Анотація

Робота присвячена алгоритму побудови моделей предметних галузей як мереж із термінів певної тематики – алгоритму формування мереж природніх ієрархій термінів. Для формування корпусу текстових документів була використана вільна доступна пошукова система, яка індексує повний текст наукових публікацій – Google Scholar. Результатом роботи стало візуальне представлення мережі термів для концепту «Cyber security».

Ключові слова: предметна область, кібербезпека, мережа термів, компактифікований граф горизонтальної видимості, мережа природніх ієрархій термінів

Вступ

Однією з ключових проблем сучасної інформаційної епохи є забезпечення інформаційної та кібернетичної безпеки. Тому питання, що стосуються кібербезпеки, з кожним днем стають актуальнішими. Це супроводжується також зростанням кількості публікацій по цій темі, зокрема, в мережі Інтернет. Величезні обсяги неструктурованих даних в Інтернеті спричинили ряд проблем, пов'язаних, в першу чергу, із пошуком в мережі необхідної інформації за певним запитом. Тож виникає потреба у розробці нових методів та підходів до структуризації даних [1].

1. Алгоритм формування мережі природніх ієрархій термінів

Одним із методів побудови термінологічних онтологій є алгоритм формування направленої мережі зі слів та словосполучень – алгоритм формування мереж природніх ієрархій термінів [2] для корпусу текстових документів. Цей алгоритм базується на використанні інформаційно–важливих елементів тексту, опорних слів та словосполучень (уніграм, біграм та триграм), методика виявлення яких представлена в роботі [3]. Алгоритм формування мереж природніх ієрархій термінів також передбачає побудову компактифікованого графу горизонтальної видимості (Compactified Horizontal Visibility Graph – CHVG) [3, 4] для термів – окремих слів, біграм та триграм, та встановленні направлених зв'язків між термами.

Як зазначено у роботі [3], алгоритм формування мереж природніх ієрархій термінів можна представити у вигляді послідовних етапів, які охоплюють попередню обробку отриманого корпусу текстових документів, виділення ключових слів та словосполучень, що є інформаційно-важливими в межах розглянутої предметної області, побудова компактифікованого графу горизонтальної видимості (Compactified Horizontal Visibility Graph — CHVG), перерахунок сортування вагових значень виділених термів за обраним ваговим критерієм та вибір із них найбільш вагомих. Кінцевим етапом є безпосереднє формування мережі природніх ієрархій термінів (з'єднання вузлів зв'язками "входження") та її відображення.

1.1. Формування корпусу текстових документів

Початковим етапом формування мережі термів, пов'язаної з певною предметною областю, є формування корпусу текстових документів. Для проведення досліджень була використана вільна доступна пошукова система, яка індексує повний текст наукових публікацій – Google Scholar (scholar.google.com). На цьому етапі було вивантажено анотації перших 592 статтей за запитом «Cyber Security».

1.2. Обробка текстових документів та виокремлення ключових термів

На етапі обробки текстових документів проведено процес попереднього лексичного аналізу – розбиття тексту на елементарні одиниці (токени або леми), вилучено стоп-слова, які не мають ніякого смислового навантаження, здійснено процес стематизації – скорочення слова до основи шляхом відкидання допоміжних частин (таких як закінчення чи суфікс) й подальше зважування й виокремлення термів.

В якості вагових значень термів, для формування часового ряду в якості функції, яка ставить у відповідність слову число, в даному дослідженні використовується статистичний показник важливості терма -– глобальний TF (Global Term Frequency, GTF), що дорівнює відношенню загальної кількості появи терма у всіх документах корпусу до загальної кількості

[а] dwlande.o@gmail.com
[б] dmytrenko.o@gmail.com

термів у документах корпусу. Використання цього показника дає змогу уникнути ситуації, що виникає під час роботи з текстовим корпусом заздалегідь визначеної тематики, коли інформаційно–важливий терм зустрічається майже у кожному документі корпусу і має низький ваговий показник TF.

1.3. Алгоритм побудови компактифікованого графу видимості

Для послідовності термів та їх вагових значень будується компактифікований граф горизонтальної видимості (CHVG). Загалом, мережа слів з використанням алгоритму горизонтальної видимості будується у три етапи. На першому етапі на горизонтальній осі відмічається ряд вузлів, кожен з яких відповідає словам у тому порядку, в якому вони з'являються в тексті, а по вертикальній осі відкладаються вагові значення – числові оцінки. На другому етапі будується граф горизонтальної видимості. Третій етап полягає в тому, що отримана на попередніх етапах мережа копмактифікується. В результаті буде отримано нову мережу слів – компактифікований граф горизонтальної видимості (CHVG).

1.4. Формування мережі природніх ієрархій термінів

Наступним кроком є перерахунок вагових значень, що відповідають термам у CHVG. Ця процедура дозволяє врахувати в подальшому також ті терми, які мають велике значення для загальної тематики текстового корпусу [3]. Під час виконання досліджень перерахунок ваг здійснюється з використанням алгоритму HITS [5, 6], завдяки якому визначається авторство чи посередництво для кожного вузла CHVG. Вибір форми вагового значення (авторство чи посередництво) немає значення, оскільки граф є ненаправленим Після цього всі терми упорядковуються за спаданням розрахованих вагових значень відповідних їм вузлів у CHVG.

Далі експертним методом визначається необхідний розмір (число N) створюваної мережі природніх ієрархій термінів, після чого вибирається N простих слів, біграм та триграм (всього N+N+N елементів), що мають найбільші значення вагових показників відповідних їм вузлів у CHVG.

На наступному етапі будується сама мережа природніх ієрархій термінів, в якій вузли відповідають відібраним термам, а зв'язки між ними -– входженням одного терма в інший.

2. Візуалізація й аналіз результатів дослідження

Як було зазначено раніше, для проведення досліджень в даній роботі використано корпус заздалегідь вибраних текстових документів, що тематично пов'язані з актуальною предметною областю – «Cyber Security».

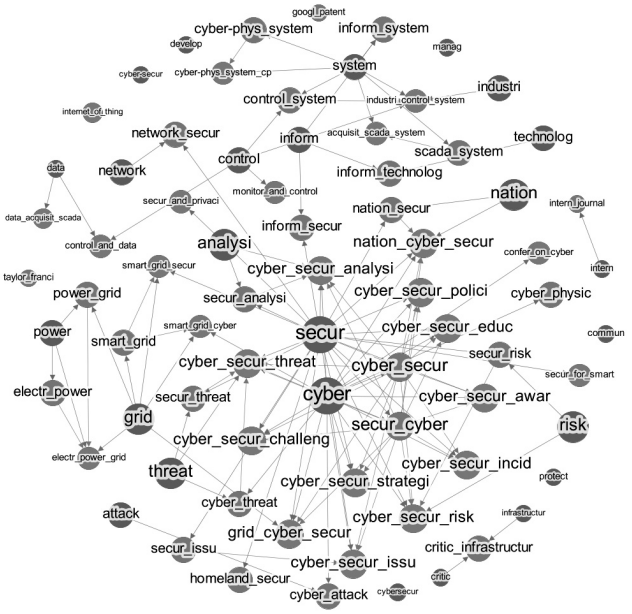

Рис. 1. Мережа природніх ієрархій термінів розміром 25+25+25 для предметної області «Cyber Security»

У таблиці 1 наведені списки найбільш вагомих термів (слів, біграм та триграм) для досліджуваної предметної області відповідно до мережевого рангового критерію HITS.

Використовуючи засоби програмного забезпечення для моделювання та візуалізації графів – Gephi (gephi.org) [7] побудована мережа природніх ієрархій термінів розміром 25+25+25 була візуалізована (Рис. 1).

Також за допомогою засобів програмного забезпечення Gephi були отримані такі параметри створеної мережі: кількість вузлів – 75; кількість зв'язків – 122; щільність мережі – 0.022; кількість зв'язаних компонент – 12; середня довжина шляху – 1; середній коефіцієнт кластеризації – 0.166. За топологічною особливістю побудована мережа має малий середній коефіцієнт кластеризації. Це пояснюється наявністю в мережі великої кількісті понять, сусіди яких мало пов'язані один з одним – це є ознакою так званих квазіієрархічних мереж. Невелика середня довжина шляху свідчить про те, що ця мережа також є "малим світом" (Small World) [8].

Висновки

У роботі розглянуто методику створення мережі зі слів та словосполучень – алгоритм формування мереж природніх ієрархій термінів шляхом масиву тематично пов'язаних текстових документів. Розглянута методика була застосована для створення мережі термів, як моделі предметної області для сфери «Cyber Security». На основі найбільшої вільно–доступної пошукової системи, яка індексує повний текст наукових публікацій – Google Scholar, був попередньо підготовлений текстовий корпус за запитом «Cyber Security» об'ємом 592 документи. Було встановлено, що побудована мережа за топологічною особливістю має малий середній коефіцієнт класте-

Табл. 1. Списки найбільш вагомих термів (слів, біграм та триграм) для «Cyber Security»

Слова	Біграми	Триграми
cyber	cyber secur	cyber secur risk
secur	smart grid	industri control system
system	cyber attack	control and data
inform	critic infrastructur	cyber secur analysi
network	inform secur	cyber secur threat
cyber-secur	control system	nation cyber secur
risk	cyber-phys system	cyber secur issu
infrastructur	power grid	cyber secur challeng
attack	network secur	grid cyber secur
critic	cyber threat	internet of thing
control	inform technolog	cyber secur awar
threat	scada system	monitor and control
develop	taylor franci	confer on cyber
industri	electr power	secur and privaci
cybersecur	secur threat	smart grid cyber
data	inform system	cyber secur incid
power	cyber physic	cyber secur strategi
commun	intern journal	acquisit scada system
analysi	secur analysi	data acquisit scada
protect	secur issu	electr power grid
grid	secur risk	smart grid secur
nation	homeland secur	cyber secur educ
intern	googl patent	cyber-phys system cp
technolog	nation secur	secur for smart
manag	secur cyber	cyber secur polici

ризації. Невелика середня довжина шляху говорить про те, що ця мережа є "малим світом" (Small World).

Отже, мережа природної ієрархії термінів, що створюється повністю автоматично, може розглядатися як основа для подальшого автоматизованого формування термінологічних онтологій за участю експертів. Також результати дослідження можуть бути використані під час створення персональних пошукових інтерфейсів користувачів інформаційно–пошукових систем, що, в свою чергу, дозволить спростити процес пошуку необхідної інформації.

Перелік використаних джерел

1. Ланде Д. В., Субач Ю. І.and Бояринова Ю. Є. LaTeX 2$_\varepsilon$в примерах. — К. : ІСЗЗІ КПІ ім. Ігоря Сікрорського, 2018. — С. 300 с. — Режим доступу: `http://dwl.kiev.ua/art/oiad/oiad.pdf`.
2. Lande D. V., Snarskii A. A., Yagunova E. V. Network of Natural Hierarchies of Terms of News Messages on Events.
3. The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text / D. V. Lande, A. A. Snarskii, E. V. Yagunova, E. Pronoza // Proceedings of the 12th Mexican International Conference on Artificial Intelligence. — 2013. — P. 209–215.
4. From time series to complex networks: the visibility graph / L. Lacasa, B. Luque, J. Ballesteros, F.and Luque, J.C. Nuño // Proc. Natl. Acad. Sci. USA 105. — 2008. — P. 4972–4975.
5. Kleinberg J. M. From time series to complex networks: the visibility graph // In Processing of ACM-SIAM Symposium on Discrete Algorithms. — 1998. — no. 46.
6. N. Langville A., D. Meyer C. Google's PageRank and beyond: the science of searchengine rankings. — 2011.
7. K. Cherven. Network Graph Analysis and Visualization with Gephi. — Packt Publishing, 2013.
8. J. Kleinberg. Navigation in a small world. — Nature, 2000.