DOI 10.31891/2307-5732-2023-319-1-338-343
УДК 004.02

СИНЬКО АННА
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-8355-461X
e-mail: anna.i.synko@lpnu.ua
ЖЕЖНИЧ ПАВЛО
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-2044-5408
e-mail: pavlo.i.zhezhnych@lpnu.ua

МЕТОД АВТОМАТИЗОВАНОГО ВИЯВЛЕННЯ ТЕРМІНІВ СТАТЕЙ ЗА ДОПОМОГОЮ ДЕРЕВА ДЛЯ ПРИЙНЯТТЯ РІШЕНЬ

З кожним днем все більше зростає кількість користувачів віртуальних спільнот, а отже і даних, що виникають під час комунікації між ними. Розміщені дані можуть містити цінне інформаційне наповнення, адже містять не тільки думку виробника, але і споживацький досвід про певний продукт. Але через те, що віртуальні спільноти мають слабку структурованість щодо подачі інформації, є більш орієнтовані на розважальний контент – можуть містити дані, які не несуть смислового навантаження, а також при розміщенні даних не всі користувачі передбачають техніки, що допоможуть збільшити релевантність пошуку цих даних. Тому пошук цільових даних потребує значних часових витрат. Для покращення пошуку даних у статі запропоновано метод, що дозволяє проаналізувати зміст розміщених дописів та виявити ключові слова з певної предметної області. Даний метод є автоматизованим та працює на основі попередньо розробленого словнику ключових фраз або регулярних виразів з ваговими коефіцієнтами приналежності до того чи іншого терміну. В результаті чого для кожного терміну будується дерево прийняття рішень, що визначає вагу терміну до змісту допису, статті. В роботі представлено обчислення ваги для одного терміну з частини допису спільноти CodeProject.

Ключові слова: *віртуальна спільнота, дерево прийняття рішень, IT-галузь, обробка великих даних, аналіз вмісту дописів.*

SYNKO ANNA, ZHEZHNYCH PAVLO
Lviv Polytechnic National University

METHOD OF AUTOMATED DETECTION OF ARTICLE TERMS USING A DECISION TREE

Every day, the number of users of virtual communities is increasing, and therefore the data that occurs during communication between them. The posted data can contain valuable information because they contain not only the manufacturer's opinion, but also consumer experience about a certain product. But, due to the fact that virtual communities have a weak structure in terms of providing information, they are more focused on entertaining content - they may contain data that do not carry a meaningful load, and also, when placing data, not all users foresee techniques that will help increase the relevance of the search for this data. Therefore, the search for target data requires significant time costs. To improve the search for data in the article, a method is proposed that allows you to analyze the content of posted posts and identify keywords from a certain subject area. This method is automated and works on the basis of a previously developed dictionary of key phrases or regular expressions with weighting coefficients of belonging to one or another term. As a result, a decision-making tree is built for each term, which determines the weight of the term to the content of the post, article.

At the same time, the level of location of the post in the discussion is taken into account, because the discussion contains a set of chronologically ordered posts. Posts placed at higher levels have a higher coefficient in the calculation. While posts are placed at lower levels - lower weighting factors. Identified key phrases before the specified term are ordered in descending order of weight. At each level of the tree, the total weight of key phrases must be equal to one. To process the data from the virtual communities, they were downloaded using the data consolidation technique. As a result, the concept of consolidated data storage was introduced, which allows collecting data from disparate sources. The paper presents the weight calculation for one term from part of the CodeProject community post.

Keywords: *virtual community, decision tree, IT industry, big data processing, analysis of the content of posts.*

Постановка проблеми

Згідно з опитуванням, яке було проведено Київським міжнародним інститутом соціології у травні 2022 року, щодня 78% українців здійснюють пошук інформації в мережі Інтернет [1]. З кожним днем все більше зростає кількість користувачів соціальних мереж, і, станом на 2022 рік їх чисельність складає 4,65 млрд, що є 58,7% від всього населення. Одним із типів віртуальних спільнот (ВС) є соціальні мережі де користувачі об'єднуються для обговорення певної тематики. Важливою рисою ВС є те, що вони містять дані основані на досвіді користувачів. Розміщені дані можуть бути корисні як для інших користувачів так і для розробників чи виробників певних продуктів. До недоліків розміщеної інформації у ВС належать:

- швидкий ріст обсягів інформації, пошук та обробка яких займе багато часових витрат;
- перевірка достовірності даних, що розміщують користувачі;
- слабка структурованість щодо подачі інформації, незважаючи на наявність тем, що групують дані довкола певних тематичних ситуацій.

Перевірка достовірності даних є розглянутою в інших наукових роботах [2, 3]. Натомість в цій статті мова йде щодо аналізу вмісту інформаційного наповнення ВС – дискусій, що містять множину дописів, що семантично пов'язані та хронологічно упорядковані [4]. Запропонований метод автоматизованого виявлення термінів дозволить визначити значущі, вагомі терміни щодо вмісту дискусії

ВС. Що є подібним на ключові слова, які задають при опублікуванні матеріалів на будь-яких веб ресурсах для підвищення релевантності пошуку даних.

Аналіз останніх джерел

У попередніх дослідженнях автори [5] провели аналіз типів спільнот та їх ознак. У результаті була побудована модель спільноти, яка показує загальну структуру, характерну для всіх типів віртуальних спільнот. Основними складовими структури ВС є учасники та інформаційне наповнення. У роботі [6] автор розробив алгоритм визначення адекватності даних інформаційного образу учасника віртуальних спільнот, що дозволяє вирішити задачу перевірки достовірності інформації користувачів, які можуть публікувати дописи. У роботі [7] наводиться оцінка дописів, яка залежить від рейтингу та способу спілкування з автором, який його опублікував. Автор [2] навів опис перевірки достовірності розміщеного інформаційного наповнення у спільнотах – текстової і мультимедійної інформації за допомогою сервісів: Findexif.com, Foto Forensics, Google Search by Image, JPEGSnoop, TinEye, Snopes.com, PeopleBrowsr, HuriSearch, Geofeedia, Verify.org,ua, Lazy Truth, Trooclick.

Питанням дослідження побудови дерев прийняття рішень наведено в роботі [8].

Ґрунтуючись на вище згаданих роботах науковців щодо побудови дерев прийняття рішення при дослідженні даних і відсутності аналізу змісту інформаційного наповнення дописів щодо ключових слів, які відображають основний зміст дискусій та покращують пошук інформації постає актуальним розроблення методу автоматизованого виявлення термінів статей/дописів.

Метою роботи є розробка методу автоматизованого виявлення термінів статей за допомогою дерева для прийняття рішень.

Виклад основного матеріалу

Для розробки методу обрано галузь інформаційних технологій (ІТ), адже це єдина галузь, що надалі продовжує зростати та розвиватися в Україні незважаючи на військові події, тим самим робить нашу країну з найбільш інноваційним розвитком по відношенню до інших країн [9].

Так як ВС є гетерогенними, різнорідними даними, тому для аналізу їх вмісту необхідно попередньо завантажувати дані за допомогою процедури інтеграції даних – консолідації, у сховище даних [11] рис. 1.

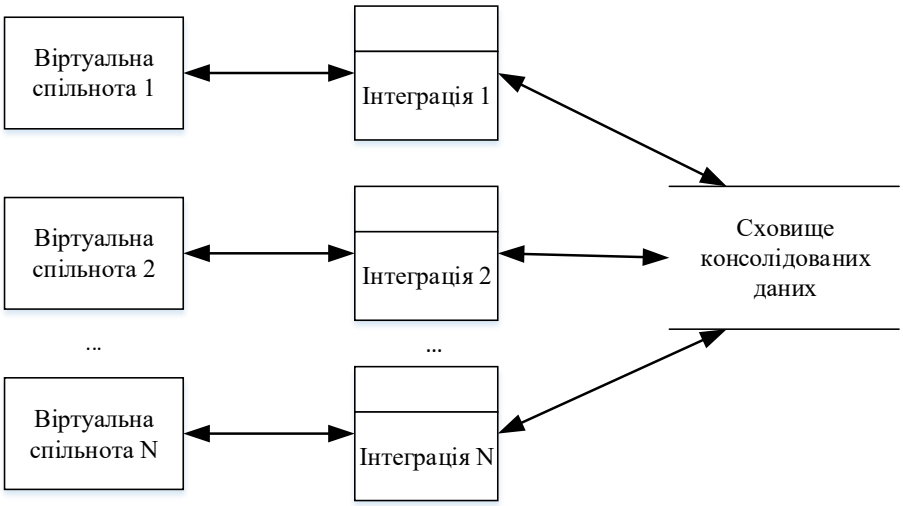

Рис. 1. Схема передачі даних з віртуальних спільнот у сховище консолідації даних

Для розмежування даних що є у сховищі консолідованих даних (СКД) та ВС введено для СКД наступні поняття: стаття, що еквівалента дискусії ВС, повідомлення – допис ВС.

Інформаційне наповнення або текст статті має ключові фрази, що зберігається у вигляді символів маски або регулярних виразів (має складну побудову), які відповідають певним термінам програмного забезпечення (ПЗ).

Термін – це здебільшого однозначне слово чи словосполучення, що виступає назвою поняття [12]. Терміни мають деревовидну структуру, де коренем виступає сам термін, а листями ключові фрази, які йому належать. Ключова фраза – слово або вислів, що містить змістовне навантаження до певного терміну. Один термін може стосуватися множини ключових фраз з певними ваговими коефіцієнтами. В свою чергу одна ключова фраза може відноситися до декількох термінів з різними ваговими коефіцієнтами. Деякі терміни можуть мати однаковий набір ключових фраз, але з різними ваговими коефіцієнтами. Всі терміни та відповідні їм ключові фрази з ваговими коефіцієнтами повинні зберігатися у словнику, який попередньо розробляє фахівець, експерт.

Схема обміну даних – виявлення термінів статті на основі ключових фраз наведено на рис. 2.

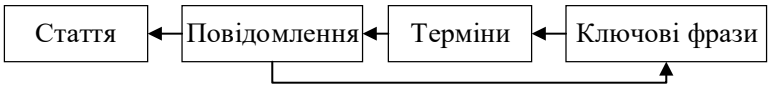

Рис. 2. Загальна схема передачі даних для виконання даного методу

Для визначення ваги термінів статті потрібно побудувати дерево для прийняття рішення. Дерево прийняття рішень є графічним методом, який пов'язує вузли прийняття рішень (терміни), можливі стратегії (ключові фрази) та наслідки їх застосування враховуючи зовнішні фактори – наявність ключових фраз які належать до терміну для обчислення міри відповідності цього терміну до статті.

Побудова дерева для прийняття рішення щодо терміну полягає в наступному. Виявлені ключові фрази до зазначено терміну необхідно впорядкувати за спаданням ваг. Вагові коефіцієнти ключової фрази до терміну наведено в словнику. На кожному рівні дерева загальна вага ключових фраз повинна дорівнювати одиниці. Обчислити значення ваги ключової фрази що не належить терміну на кожному рівні можна наступним чином:

$$\overline{w}(n) = 1 - w(n), \tag{1}$$

де *w(n)* – вага ключової фрази що належить терміну.

На основі вагових значень ключових фраз побудувати дерево, сумарна вага листків на кожному рівні дорівнюватиме одиниці (рис. 3а).

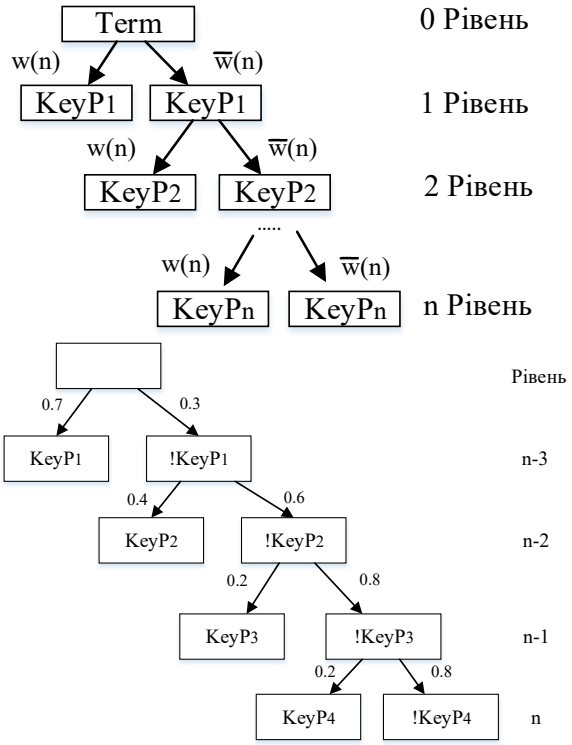

Рис. 3. Побудова дерева для визначення вагового коефіцієнту терміну: а) загальна схема; б) приклад побудови дерева рішень для терміну статті

Обчислення ваги певного терміну до тексту допису відбувається наступним чином:

$$W(Term_i) = \left((w(n) \cdot \overline{w}(n-1) + w(n-1)) \cdot \overline{w}(n-2) + w(n-2)\right) \cdot \ldots \tag{2}$$

Обчислення відбуватиметься доки, доти не доберемось до кореню дерева.

Стаття є еквівалентною дискусії ВС, а отже має ієрархічну структуру повідомлень, де повідомлення які знаходяться на вищому рівні є вагомішими ніж ті, які розташовані на нижчих рівнях. Тому при обчисленні ваги термінів доцільно враховувати рівень розташування повідомлення у дискусії наступним чином:

$$\mu_{validity}\left(Term, TextPart(Article)\right) = (Term_i) \cdot k_{ij} \tag{3}$$

де *TextPart(Article)* – текстова частина статті, k_{ij} – коефіцієнт щодо певного повідомлення у статті.

Коефіцієнт повідомлення визначається його рівнем розташування у статі:

$$k_{ij} = \frac{1}{\log_2 level(Message_m)} \quad (4)$$

де $level(Message_m)$ – рівень розташування m-го повідомлення у статті.

Слід зазначити, що обчислення за формулою 4 потрібно проводити для повідомлень, які розташовані починаючи з другого рівня, бо $\log_2 1 = 0$. Тобто для першого повідомлення і повідомлень другого рівня коефіцієнт буде дорівнювати 1, а для всіх інших коефіцієнт є меншим.

Застосування такого розподілу оцінок коефіцієнтів забезпечує покриття наступної ситуації: якщо перше повідомлення є запитанням (коли користувач цікавиться чимось), а інші є відповідями на нього. Тобто перше повідомлення може не містити багато значущої інформації. Натомість повідомлення другого рівня вже може містити потрібні, важливі дані.

Таким же чином до кожного терміну необхідно побудувати відповідні дерева та обчислити міру вагомості враховуючи коефіцієнти. Після чого буде отримано список термінів (з обчисленими мірами вагомості, значення кожної з яких знаходитиметься в інтервалі [0,1]). Надалі потрібно відібрати значущі, вагомі терміни до вмісту статті за допомогою виконання наступної умови:

$$\mu_{validity}(Term, TextPart(Article)) = \alpha \quad (5)$$

де α – показник, що визначає порогове значення для оцінки міри вагомості, значущості терміну.

Наведемо приклад побудови дерева для прийняття рішення щодо терміну на основі частини тексту допису спільноти CodeProject що став повідомленням (рис. 4).

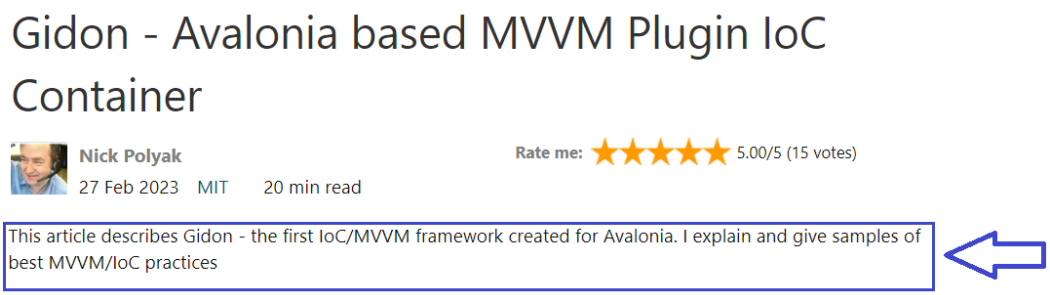

Рис. 4. Частина тексту допису спільноти CodeProject

Наведена частина тексту повідомлення, що може бути інформаційним наповненням статті містить багато ключових фраз до різних термінів. Наприклад, термін «framework MVVM» має ключові фрази з відповідними ваговими коефіцієнтами приналежності: [«Gidon»: 0.2, «first IoC»: 0.4, «MVVM framework»: 0.7, «Avalonia»: 0.2].

Надалі, згідно схеми, що представлена на рис. 3, будуємо дерево для визначення ваги терміну «framework MVVM» (рис. 3б).

Після чого, обчислюємо міру вагомості терміну до тексту повідомлення на основі виявлених ключових фраз таким чином:

$$W(Term_1) = \left((w(n) \cdot \overline{w}(n-1) + w(n-1)) \cdot \overline{w}(n-2) + w(n-2)\right) \cdot \overline{w}(n-3) + w(n-3) = 0.885 \quad (5)$$

де $w(n)$ – вага ключової фрази, що розташований на n-му рівні; $\overline{w}(n\text{-}1)$ – обчислена вага ключової фрази (за формулою 1), що розташована на рівні n-1.

Через те, що повідомлення розташоване на першому рівні тому значення коефіцієнту дорівнюватиме одиниці. При пороговому значенні α = 0,75 для оцінки міри вагомості даний термін є значущим до тексту статті *(Article(ValidTerm))* відповідно заносимо його у список термінів статті.

Отже, метод автоматизованого формування термінів певної статті полягає у наступному:

1. Для певного повідомлення статті здійснюється пошук ключових фраз в тексті на основі розробленого експертом словнику, що містить відношення між термінами та ключовими фразами з ваговими коефіцієнтами приналежності.
2. Виявлення відповідності термінів ключових фраз до термінів (слід зауважити, що кількість термінів може бути більшою за кількість ключових фраз, адже одна фраза може належати декільком термінам з різними ваговими коефіцієнтами).

3. Групування за термінами ключових фраз і побудова для кожного з них дерева для прийняття рішень (рис. 3а).
4. Обчислення вагового значення терміну (формула 3); перевірка виконання умови вагомості терміну до допису (формула 5); формування списку термінів, значення вагових коефіцієнтів яких задовольняють умові вагомості.
5. Застосування кроків 1–4 для всіх повідомлень статті.

В результаті виконання методу буде отримано список вагомих, значущих до змісту статті термінів з ваговими коефіцієнтами.

Висновки

Через те, що ВС містять дані які основані на досвіді користувачів, що можуть бути корисними як для виробників так і для споживачів певних продуктів тому пошук та аналіз розміщених даних є важливим. Для швидкого пошуку та відбору цільової інформації з ВС було наведено опис існуючих засобів, а також розроблено метод, який дозволяє проаналізувати зміст кожної дискусії ВС та визначити ключові слова / вагомі терміни. Робота методу щодо виявлення термінів здійснюється на основі попередньо розроблених експертом масок ключових фраз з ваговими коефіцієнтами приналежності до них. В результаті чого для кожного терміну будується дерево прийняття рішень і визначається його вага по відношенню до змісту дискусії. Для роботи з даними застосовано сховище консолідованих даних, а також для розмежування понять сховища і ВС введено поняття статті та повідомлення, які є еквівалентними до дискусії та допису.

Література

1. Користування інтернетом серед українців: результати телефонного опитування, проведеного 13-18 травня 2022 року. URL: https://kiis.com.ua/?lang=ukr&cat=reports&id=1115&page=12
2. Трач О.Р. Математичне та програмне забезпечення організації життєвого циклу віртуальних спільнот : дис. … к.т.н. Львів : Національний університет "Львівська політехніка", 2018. 172 с.
3. Synko A. (2022) The method of trust level of publications hosted in virtual communities. Scientific Journal of TNTU (Tern.), vol. 105, no 1, pp. 68–79. URL: https://doi.org/10.33108/visnyk_tntu2022.01.068
4. Synko Anna, Molodetska Kateryna. Application of clusterization for analysis of virtual community users. CEUR Workshop Proceedings. 2021. Vol. 2824: Proceedings of the Symposium on information technologies & applied sciences (IT&AS 2021), Bratislava, Slovak Republic, March 5, 2021. P. 9-19. URL: https://ceur-ws.org/Vol-2824/paper2.pdf
5. Пелещишин А. М., Кравець Р. Б., Сєров Ю. О. Аналіз існуючих типів віртуальних спільнот у мережі Інтернет та побудова моделі віртуальної спільноти на основі веб-форуму. Вісник Національного університету "Львівська політехніка". 2011. № 699: Інформаційні системи та мережі. С. 212-221.
6. Федушко С. С., Мельник Д. В. Розроблення алгоритму визначення адекватності даних інформаційного образу учасника віртуальних спільнот. Управління розвитком складних систем. Київ : КНУБА, 2016. № 27, p. 132–138.
7. Fedushko S., Mastykash O., Syerov Y., Shilinh A. Model of Search and Analysis of Heterogeneous User Data to Improve the Web Projects Functioning. Advances in Computer Science for Engineering and Education IV. ICCSEEA 2021. Lecture Notes on Data Engineering and Communications Technologies, Springer, Cham. 2021. № 83. P. 56-74. DOI: https://doi.org/10.1007/978-3-030-80472-5_6.
8. Сисоліна Н. П., Савеленко Г. В., Нісфоян С. С. Управління ресурсним потенціалом підприємства з використанням методу побудови дерева прийняття рішень. Вчені записки Університету «КРОК». 2019. № 4 (56), 83–88. DOI: https://doi.org/10.31732/2663-2209-2019-56-83-88.
9. Дослідження Do IT Like Ukraine: IT-індустрія зростає попри все. IT Ukraine Association. URL: https://itukraine.org.ua/it-reports-do-it-like-ukraine.html
10. Жежнич П. І. Консолідовані інформаційні ресурси баз даних та знань : навчальний посібник. Львів : Вид-во Національного університету “Львівська політехніка”, 2010. 212 с.
11. Васильченко В. Термін, термінологія. Цікаві лінгвістичні терміни. URL: https://uain.press/blogs/termin-terminologiya-tsikavi-lingvistichni-termini-923680Word

References

1. Korystuvannia internetom sered ukraintsiv: rezultaty telefonnoho opytuvannia, provedenoho 13-18 travnia 2022 roku. URL: https://kiis.com.ua/?lang=ukr&cat=reports&id=1115&page=12
2. Trach O. R. Matematychne ta prohramne zabezpechennia orhanizatsii zhyttievoho tsyklu virtualnykh spilnot: Dysertatsiia na zdobuttia naukovoho stupenia k.t.n./ Markiv Oksana Oleksandrivna. – Lviv: Lviv Polytechnic National University, 2018. – 172s.3. Fedushko S. Development of a software for computer-linguistic verification of socio-demographic profile of web-community member. Webology. – 2014. – №11(2), article 126.
3. Synko A. (2022) The method of trust level of publications hosted in virtual communities. Scientific Journal of TNTU (Tern.), vol. 105, no 1, pp. 68–79. Available at: https://doi.org/10.33108/visnyk_tntu2022.01.068
4. Synko Anna, Molodetska Kateryna. Application of clusterization for analysis of virtual community users // CEUR Workshop Proceedings. – 2021. – Vol. 2824: Proceedings of the Symposium on information technologies & applied sciences (IT&AS 2021), Bratislava, Slovak Republic, March 5, 2021. – P. 9-19. – Available at: https://ceur-ws.org/Vol-2824/paper2.pdf
5. Peleshchyshyn A. M. Analiz isnuiuchykh typiv virtualnykh spilnot u merezhi Internet ta pobudova modeli virtualnoi spilnoty na

osnovi veb-forumu / A. M. Peleshchyshyn, R. B. Kravets, Y. O. Sierov // The Journal of Lviv Polytechnic National University "Information Systems and Networks" – 2011. – Volume 699. – S. 212-221.

6. Fedushko S. S. Rozroblennia alhorytmu vyznachennia adekvatnosti danykh informatsiinoho obrazu uchasnyka virtualnykh spilnot / S. S. Fedushko, D. V. Melnyk // Management of development of difficult systems. – Kyiv: KNUBA, 2016. – № 27, pp. 132 – 138.

7. Fedushko S. Model of Search and Analysis of Heterogeneous User Data to Improve the Web Projects Functioning / S. Fedushko, O. Mastykash, Y. Syerov, A. Shilinh // Advances in Computer Science for Engineering and Education IV. ICCSEEA 2021. Lecture Notes on Data Engineering and Communications Technologies, Springer, Cham. – 2021. – № 83. – P. 56-74. – DOI: https://doi.org/10.1007/978-3-030-80472-5_6

8. Sysolina, N. P., Savelenko, H. V., & Nisfoyan, S. S. (2019). Upravlinnia resursnym potentsialom pidpryiemstva z vykorystanniam metodu pobudovy dereva pryiniattia rishen. Scientific Notes of «Krok» University, №4 (56), 83–88. DOI: https://doi.org/10.31732/2663-2209-2019-56-83-88.

9. Doslidzhennia Do IT Like Ukraine: IT-industriia zrostaie popry vse. IT Ukraine Association. URL: https://itukraine.org.ua/it-reports-do-it-like-ukraine.html

10. Zhezhnych P. I. Konsolidovani informatsiini resursy baz danykh ta znan: education manual / P.I. Zhezhnych. – Lviv: Publishing House of Lviv Polytechnic National University, 2010. – 212 s.

11. Vasylchenko V. Termin, terminolohiia. Tsikavi linhvistychni terminy. URL: https://uain.press/blogs/termin-terminologiya-tsikavi-lingvistichni-termini-923680Word