

УДК 004.67

І. В. Балагура

Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна

Визначення експертних груп на основі аналізу реферативної бази даних «Україніка наукова»

Мета статті полягає у розв'язанні завдання об'єктивного визначення експертних груп фахівців за ознаками високої кваліфікації, володіння спеціальними знаннями для здійснення наукової чи науково-технічної експертизи. Запропоновано алгоритм для здійснення вибірки експертних груп за формальними ознаками. Виявлення експертних груп може бути здійснено за допомогою визначення кластерів у мережі співавторства та побудови відповідної мережі термінів. Представлено мережі термінів для наукових колективів з інформатики.

Ключові слова: реферативна база даних «Україніка наукова», український реферативний журнал «Джерело», мережі співавторів, мережі термінів, експертні групи.

Вступ

Наука є джерелом нового знання та забезпечує розвиток основних складових держави, таких як промисловість, освіта, оборона, державна безпека та ін. Послаблення зв'язків з наукою або зниження конкурентоспроможності однієї зі складових зупиняє розвиток держави в цілому. Національний науковий розвиток має винятковий вплив на державну безпеку та оборону, адже саме ці напрями не можуть повністю ґрунтуватися на закордонних розробках. Тому актуальним завданням є підбір компетентних експертів, залучення науковців до вирішення важливих державних завдань. Ефективний розв'язок задачі швидкого і кваліфікованого визначення експертних груп сприятиме успішному вирішенню важливих завдань, які потребують залучення спеціалістів з різних галузей.

Інформатизація людської діяльності зумовила значне збільшення кількості інформації, щодня виникають нові напрями наукових досліджень, що сьогодні унеможлиблює пошук експертів за допомогою анкетування та опитування [1]. Актуальним завданням для дослідників є пошук компетентних експертів і формування експертних груп на основі інформаційних пошукових систем, соціальних мереж, наукових баз даних.

© І. В. Балагура

Основна ідея дослідження полягає у розв'язанні завдання об'єктивного визначення експертних груп фахівців за ознаками високої кваліфікації, володіння спеціальними знаннями для здійснення наукової чи науково-технічної експертизи. Запропонований алгоритм дозволить здійснити вибірки експертних груп за формальними ознаками. Закордонними вченими проведено ряд досліджень, присвячених теоретичним аспектам і вимогам для виділення експертних груп. У тому числі у роботах [2, 3] описано критерії погодження наукової і державної політики, а також промисловості для об'єктивного вибору експертів. У роботі [4] запропоновано підхід для пошуку експертів на форумах та у спеціалізованих соціальних мережах, де експерт має можливість розкрити свої знання та спосіб мислення. Основні сучасні методи визначення експертів ґрунтуються на визначенні ключових слів та інформаційному пошуку в масивах інформації відповідних особистостей. У [5] авторами запропоновано пошук експертів на основі аналізу наукових текстів і побудови онтологій предметної галузі. В роботах [6, 7] підкреслено важливість зв'язку науки, промисловості та виробництва задля прийняття важливих рішень і залучення експертів-науковців у всі галузі державної діяльності. Сучасні роботи надають описи формування спеціалізованих профілів науковців, пропонуються методи пошуку, що використовують імовірнісні моделі та інші методи аналізу тексту. Залучення науковців для прийняття важливих рішень необхідні для оборони держави, в охороні здоров'я, промисловості. Значна кількість публікацій з даної тематики вітчизняних і зарубіжних науковців зумовлює актуальність представленої тематики.

Постановка завдання

Виявлення експертних груп може бути здійснено за допомогою визначення кластерів у мережі співавторства [8, 9]. Авторами запропоновано побудувати таку мережу на основі реферативної БД «Україніка наукова», яка є найбільшим політематичним ресурсом в Україні. Сьогодні реферативна база даних «Україніка наукова» нараховує понад 571 598 записів. Даний реферативний ресурс створено в 1998 році для забезпечення вільного доступу до інформації про результати наукової діяльності вітчизняних учених і фахівців, на основі розвинення українського реферативного журналу «Джерело». Цьогоріч український реферативний журнал «Джерело» святкує своє 20-річчя [10].

У роботі [9] запропоновано методику наукометричного аналізу на основі поєднання мереж співавторів і мереж термінів. Представлений алгоритм є основою для визначення експертних груп у даній роботі. Детальний наукометричний аналіз окремого наукового напрямку чи галузі дає змогу визначити найбільш впливових науковців і тематику їхніх досліджень. Саме поєднання методів аналізу мереж співавторів і термінів для аналізу баз даних наукових публікацій надає змогу визначити найбільш компетентних експертів в окремих наукових напрямках за визначеною тематикою. Методам аналізу мереж співавторів присвячено багато зарубіжних наукових праць [11, 12]. В Україні також проводяться дослідження мереж співавторів, зокрема, О.С. Майстренко в дисертації побудував мережі співавторів для вибраних науковців із НТУУ «КПІ», а О.І. Мриглод — для вибраних наукових видань [13, 14].

Методика визначення експертних груп у своїй основі містить 4 етапи, що включають методи складних мереж, методи фільтрації тексту, візуалізацію даних.

У попередніх роботах [9] наведено методику досліджень і представлено відповідні дослідження реферативної бази даних «Україніка наукова» перших двох етапів.

На першому етапі визначається галузь і наукові напрями, за якими буде проведено аналіз, завантажуються та фільтруються файли із реферативною інформацією. Результатом першого етапу є матриця мережі — відфільтровані за певними науковими напрямками дані про авторів і зв'язки між ними [8].

Другим етапом є створення мережі співавторів досліджуваної галузі, визначення основних характеристик мережі. У ході другого етапу визначаються основні властивості співробітництва науковців, наукові групи та найбільш комунікативні науковці за певним науковим напрямком [15].

На третьому етапі відбувається відбір рефератів наукових публікацій найбільш комунікативних науковців і створення текстового корпусу для виділення основних термінів за науковими напрямками [16].

На четвертому етапі проводиться візуалізація мереж термінів науковців і галузі в цілому, розрахунок основних параметрів і виділення опорних слів і відповідних словосполучень, автоматизоване формування термінологічної основи онтології наукових напрямків та експорт даних до редакторів онтологій. Проводиться узагальнення результатів, опис основних характеристик, тенденцій у галузі, остаточне погодження експертного колективу [9].

Результатом роботи має бути перелік науковців, що можуть бути експертами в заданому напрямі та перелік основних термінів, які відповідають тематиці робіт експертів.

Результати досліджень

Відповідно до третього етапу запропонованої методики, необхідно провести відбір рефератів наукових публікацій найбільш комунікативних науковців. У дослідженні використовувалися реферати наукових колективів чотирьох провідних авторів з інформатики українською мовою, представлені в реферативній БД «Україніка наукова». Колективи авторів визначалися на основі аналізу мереж співавторів і виділення найбільш важливих вузлів за характеристиками центральності [15]. Для кожної групи науковців було сформовано текстовий корпус, що містив близько 35-ти рефератів статей у фахових журналах, тез доповідей на конференціях і монографій. При попередній обробці тексту було вилучено нетекстові символи та виконано стеммінг.

Основні терміни та термінологічні словосполучення визначалися за допомогою статистичного методу TF-IDF та методу «сигма». При подальшому обробленні для послідовності термінів і їхніх вагових значень будувалися компактифіковані графі горизонтальної видимості та виконувалися перевизначення значень вагових коефіцієнтів. Далі експертним методом і за допомогою стоп-словника обирались основні терміни, що використовує автор [16].

Наступним етапом стало представлення мережі термінів наукових колективів та їхній аналіз. Відібрані терміни використовуються для створення мережі термі-

нів, де вузлами є терміни та словосполучення, а зв'язки відповідають входженню термінів до словосполучень. На рис. 1 представлено приклад мережі термінів для наукового колективу автора академіка НАН України В.В.Петрова, створеної за допомогою програмних засобів Gephi.

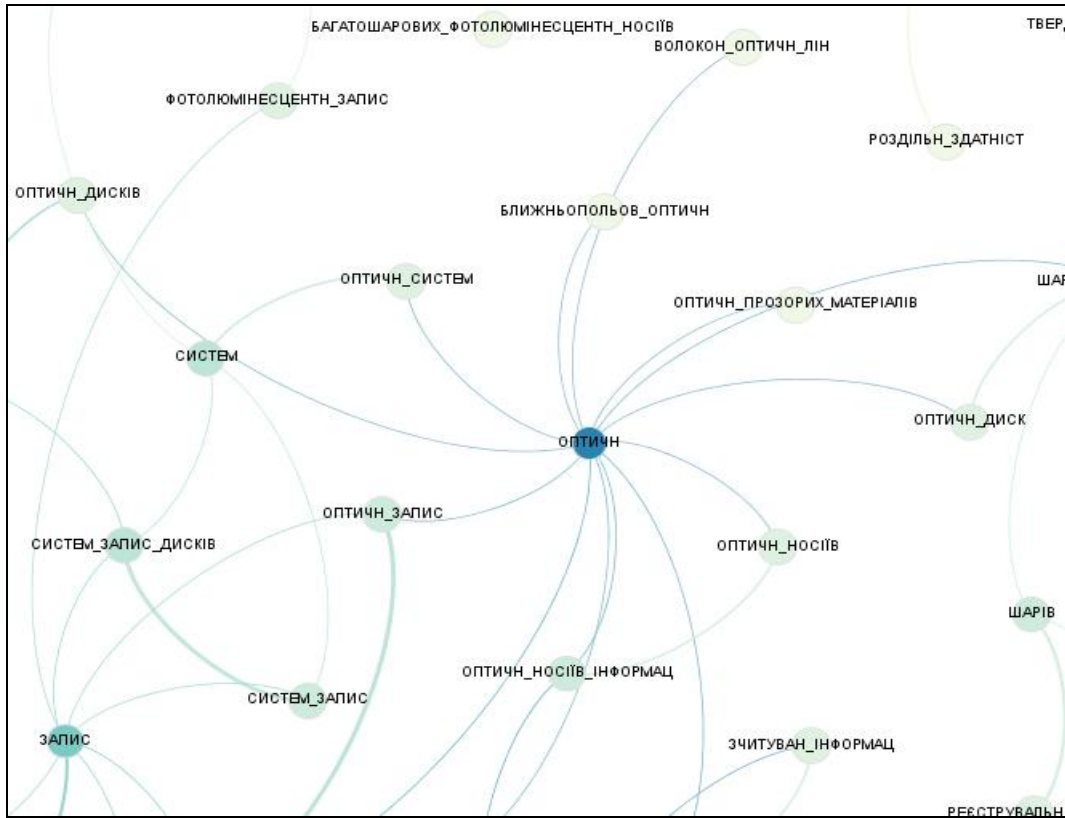


Рис. 1. Мережа термінів на основі аналізу публікацій наукового колективу В.В. Петрова

Основні терміни складаються з одного, двох та трьох слів та описуються за допомогою матриці інцидентності A , а матриці $A^T A$ та AA^T містять асоціативні зв'язки першого (якщо два терміни утворюють третій) та другого (якщо термін утворений двома іншими) роду [15]. На рис. 2 наведено мережу термінів із асоціативними зв'язками.

Найбільш значимі слова і словосполучення в мережі термінів можна виділяти за допомогою алгоритму HITS (hyperlink induced topic search). Алгоритм HITS заснований на визначенні «авторитетних вузлів» (вузлів, на які багато посилань) та «вузлів-посередників» (вузли, що найбільш часто посилаються на інші). Для наукового колективу автора В.В. Петрова найбільш значимі терміни наведені у таблиці. Таким чином, даний науковий колектив необхідно залучати при вирішенні завдань запису, зберігання, зчитування інформації, використання оптичних носіїв.

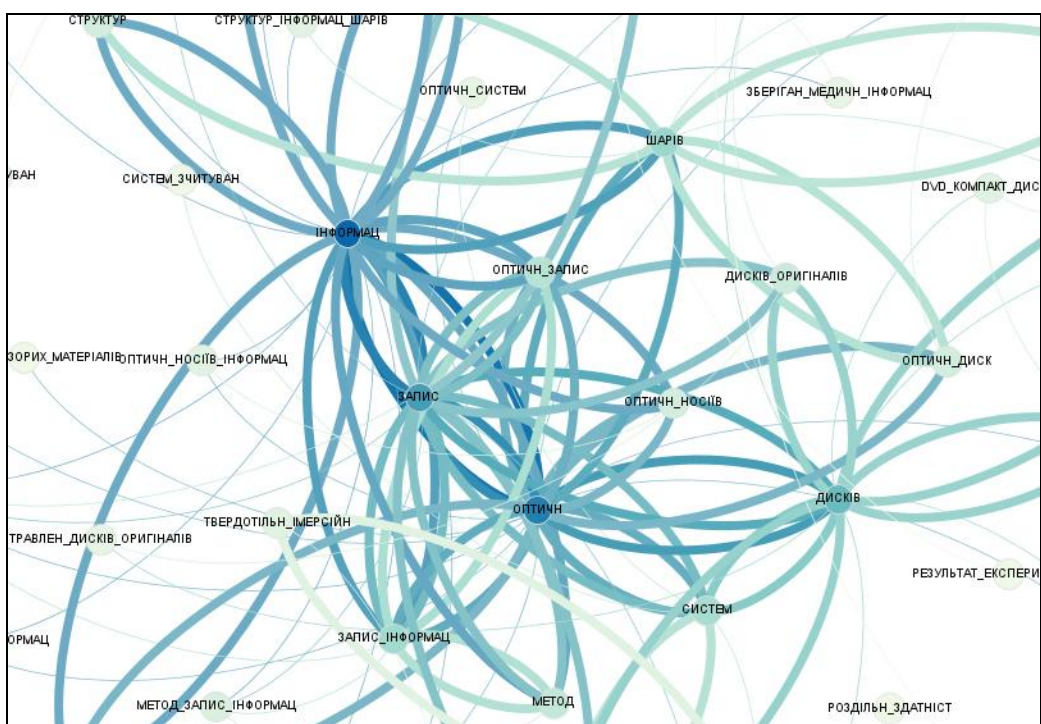


Рис. 2. Мережа термінів на основі аналізу публікацій наукового колективу В.В. Петрова

Терміни з найбільшим значенням авторства та посередництва

Терміни з найбільшим значенням авторитетності	Терміни з найбільшим значенням посередництва
оптичн_запис_інформац	інформац
оптичн_зчитуван_інформац	оптичн
оптичн_носіїв_інформац	запис
інформац_оптичн_носіях	запис_інформац
метод_запис_інформац	дисків
запис_інформац	оптичн_запис
оптичн_запис	систем
структур_інформац_шарів	зберіган_інформац
термін_зберіган_інформац	шарів
довготермінов_зберіган_інформац	зчитуван_інформац
зберіган_медичн_інформац	оптичн_носіїв
зчитуван_інформац	метод
зберіган_інформац	структур

За допомогою перетину мереж термінів окремих колективів авторів можна визначити основні терміни наукового напрямку, в якому вони працюють (рис. 3). Така мережа дає можливість визначити загальні напрями експертних груп або сформулювати колектив із окремих експертів, що охоплюють декілька напрямів досліджень.



Рис. 3. Мережа термінів одних із найбільш комунікативних науковців з комп'ютерних наук

На рис. 3 відображено наукові колективи, які для наочності названі прізвищем автора найбільш комунікативних авторів визначених початково. У результаті побудови спільної мережі термінів отримано загальний словник: «інформація», «система», «аналіз», «модель», «метод», «обробка», «структура», «управління», «процесор», «математична модель», «обробка інформації», «інформаційна система», «комп'ютерна система». Такі терміни характеризують перетин досліджень колективів авторів з інформатики, що представлені авторами В.В. Петровим, О.В. Палагіним, В.О. Хорошком і В.С. Харченком.

Отримані терміни за окремими науковими напрямами та дані про їхні зв'язки входження та асоціативні зв'язки можуть виступити термінологічною основою онтології, де терміни можуть виступати як класи онтологій, а асоціативні зв'язки — як відношення між ними. Така термінологічна основа онтології потребує подальшого експертного аналізу та розширення і дозволяє формалізувати науковий напрямок окремих наукових груп.

У результаті проведеної роботи визначено чотири наукові колективи, які можуть виступати як експертні групи для вирішення завдань, пов'язаних із комп'ютерними науками. Визначені терміни надають можливість коректно визначати галузь діяльності відповідних експертів.

Висновки

У роботі представлено дослідження мереж термінів для наукових колективів у інформатиці. На прикладі показано можливість пошуку експертних груп на основі мереж співавторів із використанням мереж термінів. Запропоновано технологію наукометричного аналізу для вдосконалення інструментарію вітчизняної реферативної бази, покращенню якості інформаційного сервісу та ефективності аналітичної діяльності.

Результати досліджень дають можливість обґрунтувати, автоматизувати та прискорити процедуру підбору компетентних експертів, що входять до складу сформованих галузевих груп. Зокрема, результати роботи дозволять проводити пошук експертів-науковців для державних завдань, що потребують швидкого реагування на події, пошук наукових колективів і шкіл для вирішення технічних завдань, зокрема в оборонній галузі.

1. *Рогущина Ю.В.* Використання організаційних онтологій для пошуку експертів у нових предметних областях / Ю.В. Рогущина, А.Я. Гладун // Проблеми програмування. — 2007. — № 1. — С. 73–83.

2. *Leverage RAF to find domain experts on research social network services: A big data analytics methodology with MapReduce framework* / Sun Jianshan, Xu Wei, Ma Jian, Sun Jiasen // *Int. J. Production Economics*. — 2015. — N 165. — P. 185–193.

3. *How experts are chosen to inform public policy: Can the process be improved?* / Silvia Rowe, Nick Alexander [et al.] // *Health Policy*. — 2013. — N 112. — P. 172–178.

4. *Finding experts in online forums for enhancing knowledge sharing and accessibility* / Chic-Ping Wei, Wen-Ben Lin [et al.] // *Computers in Human Behavior*. — 2013. — N 112. — P. 325–335.

5. *Крюков К.В.* О понятии формальной компетенции научных сотрудников / К.В. Крюков, О.П. Кузнецов, В.С. Суховеров // *Материалы III междунар. научно-техн. конф. (Минск, 21–23 февраля 2013 г.)*. — Минск: БГУИР. — 2013. — С. 143–146.

6. *Integrating expert profile, reputation and link analysis for expert finding in question-answering websites* / Duen-Ren Liu, Yu-Hsuan Chen, Wei-Chen Kao, Hsiu-Wen Wang // *Information Processing and Management*. — 2013. — N 49. — P. 312–329.

7. *An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities* / Guangyou Zhou, Jun Zhao [et al.] // *Knowledge-Based Systems*. — 2014. — N 66. — P. 136–145.

8. *Визначення потенційних експертних груп науковців в мережі співавторства з використанням методів підтримки прийняття рішень* / І.В. Горбов, С.В. Каденко, І.В. Балагура [та ін.] // *Реєстрація, зберігання і оброб. даних*. — 2013. — Т. 15, № 4. — С. 87–97.

9. *Балагура І.В.* Лінгвістичні дослідження взаємозв'язків науковців на основі аналізу реферативної бази даних «Україніка наукова» / І.В. Балагура, Д.В. Ланде // *Реєстрація, зберігання і оброб. даних*. — 2014. — Т. 16, № 3. — С. 45–53.

10. *Корпоративний проект «Система реферування української наукової літератури»* [Електронний ресурс]:[сайт]. — Режим доступу: <http://www.nbuv.gov.ua/node/522>

11. *Alireza A.* Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks / Abbasi Alireza, Hossain Liaquat, Loet Leydesdorff // *Journal of Informetrics*. — 2012. — Vol. 6, N 3. — P. 403–412.

12. *Leydesdorff L.* Eugene Garfield and algorithmic historiography: Co-Words, Co-Authors, and Journal Names / Loet Leydesdorff // *Annals of Library and Information Studies*. — 2010. — Vol. 57, N 3. — P. 248–260.

13. *Майстренко О.С.* Бізнес-моделювання слабо структурованих організаційних систем: автореф. дис. на здобуття ступеня канд. техн. наук: спец. 01.05.04 / О.С. Майстренко. — К.: НТУУ «КПІ», 2013. — 24 с.

14. *Мриглод О.І.* Елементи кількісного аналізу наукових періодичних видань / О.І. Мриглод // *Наука України у світовому інформаційному просторі*. — 2014. — Вип. 10. — С. 19–28.

15. *Балагура І.В.* Дослідження параметрів важливості вузлів в мережах співавторів / І.В. Балагура, Д.В. Ланде, І.В. Горбов // *Реєстрація, зберігання і оброб. даних*. — 2013. — Т. 15, № 1. — С. 45–52.

16. *Ландэ Д.В.* Подход к созданию терминологических онтологий / Д.В. Ландэ, А.А. Снарский // *Онтология проектирования*. — 2014. — Т. 12, № 2. — С. 83–92.

Надійшла до редакції 12.09.2015