

Міністерство освіти і науки України

Тернопільський національний технічний університет імені Івана Пулюя

**АНДРУЩЕНКО Валентина Борисівна**

УДК 004.031.42

**ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ НАУКОМЕТРИЧНОГО АНАЛІЗУ НА  
ОСНОВІ МОНІТОРИНГУ РЕСУРСІВ МЕРЕЖІ ІНТЕРНЕТ**

05.13.06 – Інформаційні технології

**АВТОРЕФЕРАТ**

дисертації на здобуття наукового ступеня  
кандидата технічних наук

Тернопіль – 2019

**Дисертацією є рукопис.**

**Робота виконана в Інституті проблем реєстрації інформації НАН України**

|                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|-------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Науковий керівник | доктор технічних наук, старший науковий співробітник<br><b>ЛАНДЕ Дмитро Володимирович</b><br>Завідувач відділу спеціалізованих засобів моделювання<br>Інституту проблем реєстрації інформації НАН України                                                                                                                                                                                                                                                          |
| Офіційні опоненти | доктор технічних наук, професор<br><b>ВИШНІВСЬКИЙ Віктор Вікторович</b><br>Завідувач кафедри комп'ютерних наук<br>Державного університету телекомунікацій МОН України<br><br>Доктор технічних наук, доцент<br><b>СУБАЧ Ігор Юрійович</b><br>Завідувач спеціальної кафедри №5<br>Інституту спеціального зв'язку та захисту інформації<br>Національного технічного університету України<br>«Київський політехнічний інститут імені Ігоря<br>Сікорського» МОН України |

Захист відбудеться 04 квітня 2019р. о 14 годині на засіданні спеціалізованої вченої ради K58.052.06 в Тернопільському національному технічному університеті імені Івана Пулюя за адресою: 46001, м. Тернопіль, вул. Руська, 56, ауд. 79.

З дисертацією можна ознайомитися у бібліотеці Тернопільського національного технічного університету імені Івана Пулюя за адресою: 46001, м. Тернопіль, вул. Руська, 56.

Автореферат розісланий \_\_ березня 2019 р.

Учений секретар

спеціалізованої вченої ради



Фриз М.Є.

## ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

**Обґрунтування вибору теми дослідження.** На сьогодні наукометрична інформація все більше стає інструментом для застосування інформаційних технологій, наприклад, для оцінки наукової впливовості, хоча до цього здебільшого була зосереджена тільки на ранжуванні вчених, дослідницьких груп та організацій за рахунок кількісних показників, чого, в свою чергу недостатньо для надання повноцінної картини результативності діяльності вченого чи інститутції. Дослідження за даною тематикою провадилися Ніком Хассеном і Клаудією Лоебекке. Саме розвиток і використання наукометричного інструментарію інформаційними технологіями дозволить сформувати всебічний погляд на аналіз публікаційної активності і можливість подальшого використання отриманих даних для прийняття важливих управлінських рішень.

Робота, яка провадиться з врахуванням і використанням додатків для ранжування чи інших кількісних обчислень, базується на інформації, що переважно міститься в мережі Інтернет. З огляду на те, що кожен відповідний інформаційний ресурс, що містить інформацію про науковий журнал, публікації, авторів, тощо, відрізняється один від одного не тільки дизайном (організація меню, ілюстративний матеріал, тощо) і доступом до інформації – відкритий чи передплатний, а й наборами даних, що містить ресурс та способами пошуку та представлення інформації, можливостями збереження отриманих в результаті пошуку даних, тощо. В дисертаційній роботі представлено актуальні підходи для формування необхідних критеріїв шляхом аналізу різноманітних ресурсів наукової та наукометричної інформації. Нажаль, на сьогодні більшість наукових робіт щодо формування та оцінки роботи інформаційних систем зосереджені на аналізі освітніх ресурсів та питанні захисту ресурсів наукової інформації.

Як наслідок розвитку ресурсів наукової інформації виникли нові можливості оцінки наукової інформації та дослідження закономірностей наукової взаємодії. Одним із основних інструментів вивчення наукової закономірності – є мережі співавторів, за рахунок побудови і аналізу яких можна: виокремити низку колаборацій, до яких можна долучитися при плануванні реалізації складних наукових задач, сформувати пул експертів для зовнішньої оцінки наукових рішень, проєктів, рецензування наукових публікаційних доробків. Один із провідних сервісів наукової інформації, що є не тільки пошуковим інструментом наукової літератури, а й ресурсом, що містить інформацію про авторів публікацій - Google Scholar. Запропонована у дисертаційному дослідженні методика передбачає побудову мереж співавторів – моделей співробітництва вчених за рахунок зондування ресурсу Google Scholar Citations із використанням мережі понять – тегів ресурсу Google Scholar Citations. Попередні дослідження щодо побудови мереж співавторів, їх вплив на цитування, предикативної характеристики відповідних моделей за останні 5 років представлені в роботах С. Кумара, К. Біскарро, К. Гіпонні, К. Ю, Ц. Лонг, К. Шульца, А. Мазлумена та інших. Серед вітчизняних вчених питаннями побудови мережі співавторів, їхньої оцінки та аналізу присвячені роботи Д.В. Ланде, І.В. Балагури, Ю. Головача, О. Мриглод та інших.

При проведенні наукометричного аналізу актуальним є питання не тільки взаємодії окремих учасників наукового процесу, а й взаємозв'язок між науковими напрямками та предметними областями з огляду на актуальність міждисциплінарної складової сучасних наукових процесів. За рахунок використання інформаційних технологій можлива побудова мережі предметних областей на базі інформаційних ресурсів наукової інформації, адже одним із принципів сортування та характеристик, що визначають ту чи іншу публікацію або наукові дані, є приналежність до конкретної предметної області та наукового напрямку.

Із використанням засобів математичної лінгвістики, зокрема інструментів для обробки текстових масивів можна реалізувати алгоритми сканування ресурсів наукової інформації для виокремлення інформації про науковий напрямок та співставити її із словниками, що містять повну інформацію про предметну область та відповідні наукові напрямки, і за результатами обробки отриманої інформації побудувати та візуалізувати мережу предметних областей. Така мережа дозволить актуалізувати не тільки зв'язки фахівців-науковців, як у випадку мережі співавторів, так і наукові галузі, що є важливим інструментом виокремлення міждисциплінарної складової ключових аспектів досліджень. Тому в роботі представлено моделі, відповідні алгоритми та їх програмна реалізація на базі наукометричного пошукового ресурсу Google Scholar Citations, на базі он-лайн архіву препринтів arXiv. Саме можливість представити результати із використанням різних джерел інформації дозволить не тільки розширити межі сприйняття того чи іншого концепту, для якого формується мережа предметних областей, а й обмежити застосування концепту для конкретного випадку низкою притаманних предметних областей.

Також у роботі представлено підходи до обрахунку нового наукометричного індексу, який поряд з такими звичними наукометричними показниками, як кількість публікацій, кількість посилань, індекс Гірша, зображений у роботі показник визначає популярність автора (науковця) в межах статей он-лайн енциклопедії Вікіпедія. Якщо індекс Гірша науковця дорівнює деякому числу  $h$  і науковець є автором  $h$  публікацій, кожна з яких була процитована щонайменше  $h$  разів і визначає впливовість науковця в науковому середовищі з огляду на його публікаційну активність та цитованість, то Вікі-індекс визначає впливовість науковця у сфері популярної науки, доступної широкому загалу користувачів мережі Інтернет. Інструмент для визначення індексу популярності автора можна використовувати при формуванні стратегії спрямованої на збільшення видимості науковця, автора, в межах он-лайн енциклопедії Вікіпедія і таким чином дозволить дати відповідь на питання щодо необхідності більшої присутності науковців, і відповідно наукових груп, інституцій в енциклопедичному ресурсі та популяризації дослідження.

Запропоновані в роботі моделі та підходи розширюють межі використання інформаційних технологій для проведення наукометричного аналізу та формування масивів даних на базі он-лайн ресурсів наукової та наукометричної інформації. Використання розроблених у роботі підходів дозволить науковцям: скоротити час на пошук та виокремлення необхідної інформації, отримати

інструменти для прийняття рішень щодо визначення експертних груп для оцінки наукових проєктів, результатів реалізації наукових досліджень, рецензування наукових публікацій; отримати новий актуальний інструмент для пошуку партнерів для реалізації великих наукових проєктів в рамках дослідницьких колаборацій поряд із звичними підходами і способи формування стратегії популяризації досліджень і доведення їх до широкого загалу споживачів наукової інформації.

**Зв'язок роботи з науковими програмами, планами, темами, грантами.**

Результати дисертаційної роботи та запропоновані у роботі методи, моделі та відповідні алгоритми відображено у науковому звіті Інституту проблем реєстрації інформації НАН України, відповідно до плану науково-дослідної роботи «Розробити та дослідити моделі предметних областей при формуванні баз знань і забезпеченні семантичного пошуку» (2016 рік, номер держреєстрації 016U000507), у річних звітах Державного фонду фундаментальних досліджень (2015, 2016, 2017 роки). Також запропоновано способи реалізації робіт в рамках проєкту Рамкової програми Європейського Союзу з досліджень та інновацій «Горизонт-2020» - Responsible Research And Innovation Networked Globally (RRING) (грантова угода № 788503).

**Мета і завдання дослідження.** Метою дисертаційної роботи є розробка інформаційних технологій для проведення наукометричного аналізу на основі моніторингу ресурсів мережі Інтернет, формування масивів інформації, розробки відповідних моделей і алгоритмів реалізації додаткових можливостей для існуючих наукометричних ресурсів та ресурсів наукової інформації відкритого доступу.

Для досягнення цієї мети були поставлені такі завдання:

1. розроблення моделей та алгоритмів для створення інформаційних масивів на базі ресурсу Google Scholar для побудови мереж предметних областей та співавторів;
2. визначення індексу популярності автора (науковця) в енциклопедичному ресурсі Wikipedia для збільшення обсягу пошуку в цьому ресурсі;
3. розроблення алгоритму побудови онтології поняття на базі енциклопедичного ресурсу Wikipedia;
4. забезпечення повноти формування мережі предметних областей для поняття на базі ресурсу препринтів ArXiv за моделлю «Концепт – масив наукових публікацій» та побудова алгоритму цього формування.

**Об'єктом дослідження** є аналіз ресурсів наукометричної та наукової інформації.

**Предметом дослідження** є методи і засоби інформаційних технологій побудови нових інформаційних масивів на базі ресурсів наукової та наукометричної інформації засобами теорії складних мереж і математичної лінгвістики.

**Методи дослідження.** Для реалізації поставлених задач було використано такі наукові методи: методи аналізу текстових масивів, методи статистичного аналізу, методи математичної лінгвістики, кількісні методи наукометричного аналізу, теорію графів та теорію складних мереж.

### **Наукова новизна отриманих результатів:**

- вперше розроблено моделі та алгоритми побудови предметних областей на базі ресурсу Google Scholar Citations, що забезпечило повноту отримання інформації користувачами ресурсу при оформленні власного профілю;
- вперше запропоновано критерії аналізу наукометричних ресурсів за показниками повноти та доступності інформації для користувача;
- вперше запропоновано і реалізовано на основі енциклопедичного ресурсу Вікіпедія алгоритми: а) обчислення індексу популярності автора (науковця), б) побудови онтології поняття.

**Практичне значення отриманих результатів** полягає у можливості використання реалізованих додатків, що можуть виступати додатковим інструментарієм до традиційних наукометричних показників при аналізі результативності вчених, колективів науковців, наукових установ та результативності реалізації конкурсних проектів. Запропоновані моделі дозволять виокремити міждисциплінарну складову, що розширить межі сприйняття визначеного поняття, з точки зору приналежності до певної предметної області, для залучення фахівців різних галузей знань для реалізації глобальних проектів. Важливою є можливість використання отриманих результатів з точки зору охоплення даних не тільки ресурсів звичних і доступних лише науковому загалу, а й тих, що використовуються широким колом користувачів глобальних мереж. Результати дослідження було використано при формуванні стратегії підготовки грантового дослідження Підприємством Української академії наук «Інститут системних досліджень та інформаційних технологій» та при формуванні матеріалів щодо розвитку навичок студентів та викладачів кафедри прикладної математики НТУУ «КПІ імені Ігоря Сікорського» основним елементом підготовки грантових запитів у межах Рамкової програми ЄС з досліджень та інновацій «Горизонт 2020».

### **Особистий внесок здобувача**

Безпосередньо автором здійснено:

- інформаційний пошук та аналіз літературних даних за темою дисертації;
- розробка критеріїв для оцінки ресурсів наукової та наукометричної інформації;
- формулювання поняття мережі предметних областей;
- розробка моделі «концепт – масив наукових публікацій», відповідних алгоритмів та їх програмна реалізація на базі ресурсу препринтів arXiv.

Наукові роботи опубліковані у співавторстві з Ланде Д.В., Гриньовим Б.В., Кияком Б.Р., Балагурою І.В., Красовскою О.В., Величко І.Г.

Співавторами наукових праць є науковий керівник та науковці, спільно з якими проведені дослідження.

У роботах виконаних у співавторстві, здобувачеві належить: [1, 2] – розробка моделей та нових індексів; [5, 6, 9] – розробка критеріїв для оцінки результативності фундаментальних досліджень, що ґрунтуються на наукометричних показниках учених, що враховують самоцитування; [7, 11] –

розробка підходу та елементів алгоритму побудови мережі співавторів на базі ресурсу Google Scholar Citations; [8] – оцінка наукометричної складової за результатами конкурсу спільних проєктів; [12] – розробка алгоритму та програмна реалізація побудови онтологій на базі ресурсу Wikipedia; [13] – аналіз та інтерпретація даних за результатами сканування наукометричного ресурсу Scopus; [15, 16, 24] – розробка підходів до використання інформації, що міститься у профілі користувача системи Google Scholar Citations для побудови алгоритму мережі предметної області; [17] – систематизація та опис ресурсів наукової інформації; [1, 18, 19, 22, 25] – способи використання бібліометричних даних, що містить 2-енциклопедичний ресурс Wikipedia для формування індексу популярності науковця (автора); [21] – аналіз публікацій на зазначеним пошуковим терміном в наукометричній системі Scopus, опис підходів до інтерпретації отриманих результатів, аналіз публікаційної активності науковців України за даним напрямком, [2, 3, 23] – формування поняття мережі предметних областей, розробка та реалізація моделі для побудови мережі предметних областей на базі ресурсу архів.

Постановка мети та завдань, обговорення Постановка мети та завдань, обговорення результатів проведені разом з науковим керівником дисертаційної роботи.

**Апробація матеріалів дисертації.** Результати дисертаційної роботи доповідалися та обговорювалися на Міжнародних науково-технічних конференціях «GRANT-2015» (2015, Київ) «Відкриті семантичні технології для інтелектуальних систем - OSTIS» (2016, 2017, 2018 Республіка Білорусь, Мінськ), Міжнародній науково-практичній конференції «Наукова комунікація в цифрову епоху» (2014, 2015, 2016, 2018, Київ), 18-й Міжнародній конференції «Системний аналіз та інформаційні технології - SAIT» (2016, Київ), 8-й Науково-практичній конференції «Наука України у світовому інформаційному просторі» (2016, Київ), щорічних підсумкових конференціях «Реєстрація, зберігання та обробка даних» (2016, 2017, Київ), IX та X Міжнародних науково-технічних конференціях «Інтелектуальні технології лінгвістичного аналізу» (2016, 2017, Київ), Міжнародних науково-практичних конференціях «Інформаційні технології та безпека ІТБ» (2016, 2017, Київ), Міжнародній конференції «IEEE International Conference on Data Stream Mining and Processing» (2018, Львів).

**Публікації.** Основні положення та результати дисертаційного дослідження викладено у 7 наукових публікаціях в реферованих наукових виданнях [1-7], серед яких 5 статей у наукових фахових виданнях України з технічних наук (з них 2 у виданнях, які включено до міжнародних наукометричних баз), у 2 розділах книг [8-9], 14 публікаціях в матеріалах конференцій [10-23], 2 публікаціях в архіві препринтів [24-25].

**Структура та обсяг дисертації.** Дисертаційна робота викладена на 177 сторінках машинописного тексту, складається зі вступу, 4 розділів, загальних висновків, списку використаних джерел та 6 додатків. Обсяг основного тексту дисертації складає 118 сторінок друкованого тексту. Робота ілюстрована 5 таблицями, 53 рисунками. Список використаних джерел містить 118 найменувань.

## ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обґрунтовано актуальність теми дослідження, сформульовано об'єкт та предмет дослідження, викладено мету і задачі дослідження, визначено наукову новизну і практичне значення отриманих результатів, наведено інформацію про апробацію та публікації за темою дослідження.

У першому розділі – «Джерела наукової та наукометричної інформації» - проведено огляд літературних джерел, пов'язаних із темою дослідження. Проведено аналіз основних аспектів, що відображені у науковій літературі щодо наукової інформації та ресурсів наукової та наукометричної інформації.

У розділі описано засади, що ілюструють перелік питань, які цікавлять дослідників щодо можливостей обробки інформаційних масивів, що представлені ресурсами відкритого доступу та передплачених ресурсів.

Також представлено модель користувача інформаційної системи наукової інформації (рисунок 1).



Рисунок 1 – Модель користувача при роботі з ресурсом наукової інформації

На основі моделей було запропоновано низку критеріїв щодо оцінки ресурсів наукової інформації та представлено запропоновані критерії їх у вигляді множини змінних:

$$K = \{k_1, k_2, k_3, \dots, k_n\},$$

і відповідно – множину ресурсів, що оцінюються:

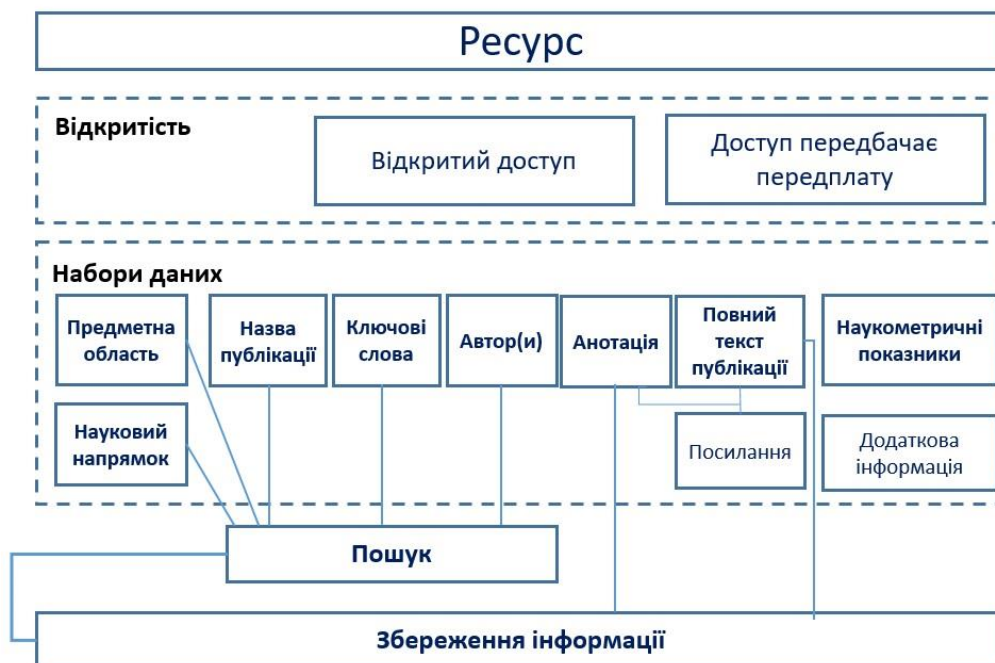
$$S = \{s_1, s_2, s_3, \dots, s_m\}.$$

Результати формалізовані у вигляді логічної матриці, що має значення «0» та «1».

Зокрема, виділені такі питання дослідження ресурсів наукової та наукометричної інформації: порівняння систем PubMed, Scopus, Web of Science та Google Scholar, переваги та недоліки системи Google Scholar, порівняння індексу Гірша, що обраховується системами Web of Science, Scopus та Google Scholar, розгляд Google Scholar як зручного альтернативного іншим наукометричним ресурсам, що представляють інформацію про наукові публікації, наукометричні показники та підкреслюють, що відкритість системи сприяє демократизації аналізу публікацій і робить аналітику щодо публікаційної активності співробітників різних

університетів доступною і актуальною, незважаючи на їхні бюджети і спроможність передплачувати доступ до наукометричних ресурсів, порівняння цитувань в журналах з медицини.

У другому розділі – «Підходи до формування критеріїв оцінки ресурсів наукової та наукометричної інформації» — представлено описову модель інформаційного ресурсу наукової інформації (рисунк 2).



Рисунк 2 – Описова модель ресурсу наукової інформації

Також проаналізовано масив ресурсів наукової інформації, здійснено огляд сучасних реферативних, наукометричних ресурсів, репозитаріїв, архівів препринтів. За рахунок проведеного аналізу було сформовано узагальнений результат, що може виступити інструментарієм при розробці систем наукової інформації та наукометричних ресурсів.

Основну увагу було приділено ресурсам, на базі яких було реалізовано розробку додаткових інструментів для розширення можливостей і формування даних для їх подальшого використання та інтерпретації.

У третьому розділі «Основні методи та технології реалізації задач розширення можливостей інформаційних систем для наукометричного аналізу» описано основні методи, що були використані при реалізації дослідження та показано підходи до реалізації задач, що поставлені у дисертаційній роботі.

Серед методів, що були застосовані при формуванні моделей та реалізації поставлених в роботі задач: методи комп'ютерної лінгвістики, методи статистичного аналізу даних, кількісні методи наукометричного аналізу, теорія графів та теорія складних мереж.

Відповідно до окреслених завдань було проведено графеметричний аналіз тексту, який містився на веб-сторінках ресурсів – виокремлення окремих слів та словосполучень, що відповідають:

- Для системи Google Scholar - лексичним одиницям, що зображують теги - наукові інтереси – наукові напрямки користувача-науковця, які науковець зазначає самостійно у системі при реєстрації та оновленні профілю. Також було сформовано тезауруси, що містили інформацію про наукові напрямки.
- Для он-лайн енциклопедії Вікіпедія було опрацьовано корпуси текстів і виокремлено в текстових масивах власні імена та гіперпосилання для подальшої обробки отриманої інформації і її візуалізації;
- Для ресурсу препринтів arXiv було проведено складання словників наукових напрямків для кожної з предметних областей, за якими систематизовано публікації. За рахунок моніторингу результатів пошуку було проведено виокремлення назв наукових напрямків, що відповідають публікаціям, які було отримано за результатами пошуку для заданого концепту.

Також для кожного з концептів, за яким відбувався пошук було обраховано частоту, з якою заданий термін зустрічається у результатах пошуку і в той же час – обраховано інверсну частоту для наукових напрямків, що зустрічаються найбільше для заданого концепту.

Частота, з якою концепт зустрічався серед результатів пошуку CF (concept frequency), було обраховано як співвідношення кількості разів  $k_i$ , скільки даний концепт зустрічається в документі, до загальної кількості слів в документі  $n$ .

$$CF = \frac{k_i}{n} \quad (1)$$

Інверсна частота для наукових напрямків – це є логарифм відношення числа всіх документів (публікацій)  $D$  до числа документів, що містять заданий концепт –  $c$ .

$$IDF = \lg \frac{D}{c}, \quad (2)$$

У роботі було застосовано багатовимірні статистичні методи, що передбачали оцінку групи елементів – ресурсів наукової інформації за декількома запропонованими параметрами. Завдяки проведенню аналізу було визначено взаємозв'язки – спільне і відмінне в роботі систем з точки зору надання користувачеві наукової інформації. Також статистичні методи було застосовано при підрахунку кількості публікації для формування представлення обсягу публікацій відповідно до наукових напрямків в системі arXiv, та відповідно – кількості науковців, що розміщують інформацію про себе та про свої публікаційні доробки на ресурсі Google Scholar.

У рамках дисертаційного дослідження було реалізовано алгоритм побудови мережі співавторів на базі ресурсу Google Scholar Citations. Виокремлення груп співавторів є інструментом, що може бути допоміжним при побудові складних наукових колаборацій, визначенні експертів для вирішення чи оцінки різних задач. Задачу було розв'язано за рахунок використання теорії складних мереж та для побудованої мережі було обчислено найкоротший шлях, що представляє собою мінімальну відстань між вузлами і обраховується як середня за всіма парами вузлів мінімальна відстань між ними:

$$l = \frac{2}{n(n+1)} \sum_{i \leq j} l_{ij}, \quad (3)$$

де  $n$  – кількість вузлів,  $l_y$  – найкоротша відстань між вузлами  $i$  та  $j$ .

Також на базі ресурсу Google Scholar Citations було вирішено задачу побудови мережі предметних областей. Мережа розглядається як модель предметної області. Вузли даної мережі – відповідають поняттям (заданим текстам), а зв'язки – семантичному зв'язку між ними.

На малюнку 3 зображено фрагмент мережі предметних областей.

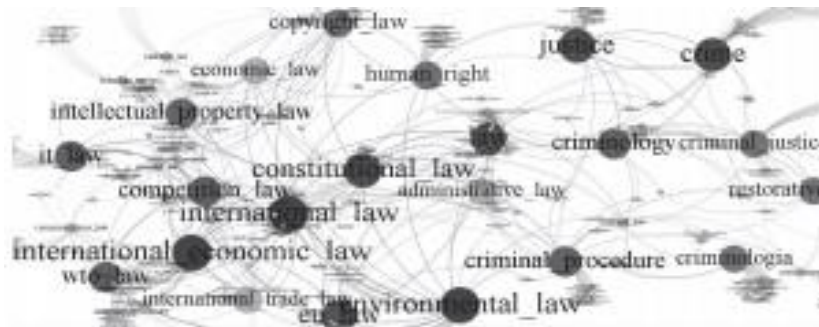


Рисунок 3 – Фрагмент мережі предметних областей за базовими поняттями.

Також у рамках дисертаційної роботи було запропоновано та реалізовано алгоритм побудови онтології поняття на базі енциклопедичного ресурсу відкритого доступу Вікіпедія. Ресурс було обрано для реалізації дослідження з огляду на відкритість, можливість внесення редагувань до статей та широкого кола користувачів.

Алгоритм передбачає врахування гіперпосилань на статті з тих чи інших понять при побудові мережі (рисунок 4).

Також було реалізовано алгоритм побудови мережі предметних областей на базі ресурсу препринтів arXiv.

З огляду на те, що на сьогодні arXiv є одним із найбільших подібних архівів і містить підготовлені до друку та обліковані матеріали за різними науковими напрямками, а коло галузей регулярно розширюється, за рахунок використання цих даних можна скласти картину застосування визначеного поняття у різних наукових галузях.

Алгоритм реалізації поставленої задачі сформовано на базі запропонованої в дослідженні моделі «Концепт – система наукових напрямків» (рисунок 5).

Модель «Концепт – система наукових напрямків» можна представити у вигляді схеми, що зображує зв'язок концепта та наукових публікацій, які містять цей концепт і масив наукових напрямків і піднапрямків, що формують словник і безпосередньо зв'язок концепту із цими параметрами.

За рахунок приналежності публікації до визначеного наукового піднапрямку та наукового напрямку визначається зв'язок концепту із науковим напрямком (предметною областю). В той же час один і той же концепт може бути відображений у публікаціях, що відносяться до різних наукових напрямків, що в свою чергу зображує мультидисциплінарну складову концепту, що задається для пошуку і подальшої побудови мережі предметних областей.

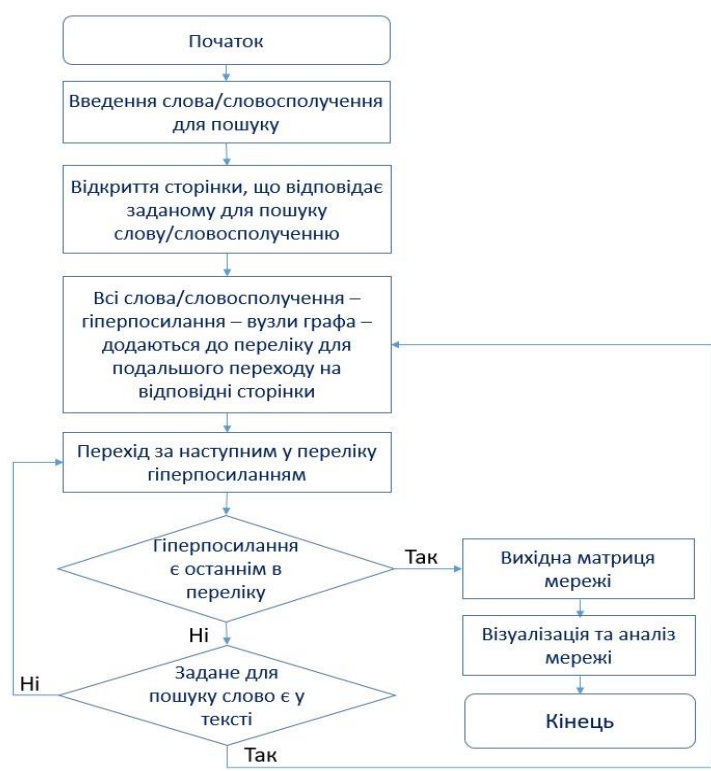


Рисунок 4 – Узагальнений алгоритм побудови онтології поняття на базі он-лайн енциклопедії Вікіпедія

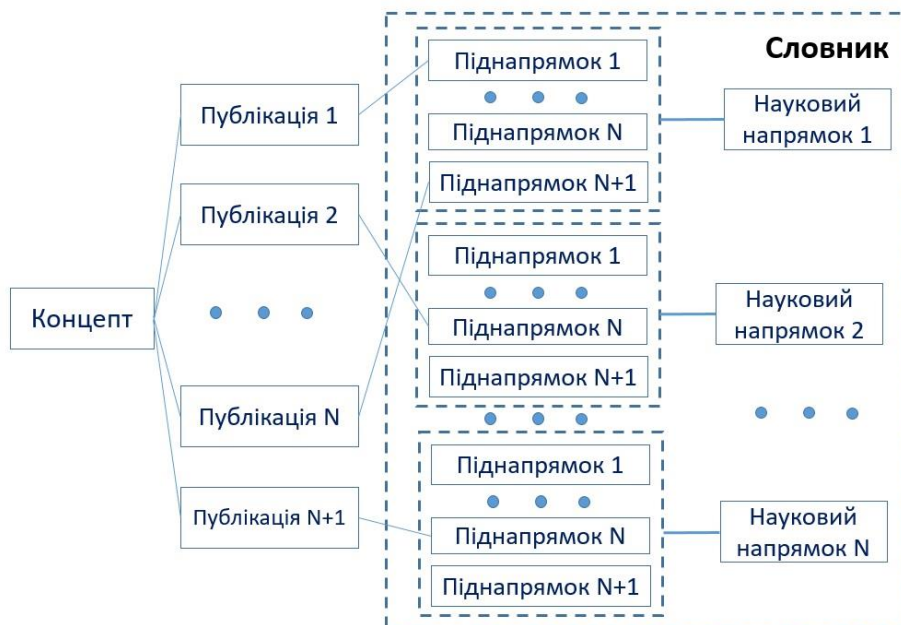


Рисунок 5 – Модель «Концепт-система наукових напрямків»

Крім візуального представлення результатів, алгоритм передбачає підрахунок кількості публікацій, що можна віднести за співставленням із запропонованим словником, до однієї предметної області, тим самим інтерпретувати результати і формувати висновки щодо притаманності заданого поняття конкретному науковому напрямку або його вагому міждисциплінарну складову.

Загальний алгоритм побудови мережі, що охоплює предметні галузі, до яких застосовується задане поняття зображено на рисунку 6.



Рисунок 6 – Загальний алгоритм побудови мережі предметних областей

На Рисунку 7 наведено приклад візуалізації результатів побудови мережі предметних областей на базі архіву препринтів arXiv.

Таким чином, за рахунок реалізації розроблених у рамках роботи алгоритмів можна побудувати нові масиви інформації за рахунок використання наукометричної інформації, що надається ресурсами відкритого доступу – енциклопедичними, наукометричними та архівом препринтів.

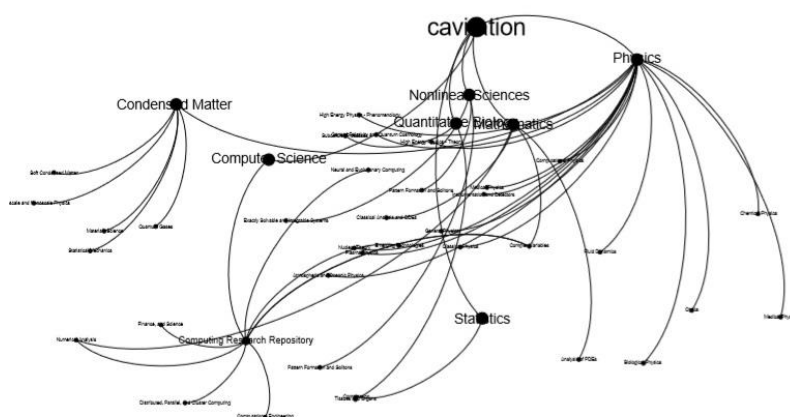


Рисунок 7 – Мережа предметних областей для заданого поняття.

У четвертому розділі – «Практична реалізація алгоритмів розширення можливостей інформаційних систем на базі наукометричної інформації» - викладено приклади практичної реалізації запропонованих технологій.

Можливості застосування запропонованих технологій було реалізовано на таких прикладах.

1. Мережа співавторів була побудована для таких базових понять:  
text mining, language processing, information retrieval, complex network.

У цьому випадку наведений алгоритм було реалізовано для групи із 10 вчених, які за даними Google Scholar Citations мають найбільшу кількість цитувань.

Під науковою школою розуміється неформальний творчий колектив учених різних поколінь. Їх об'єднує спільний стиль проведення досліджень, що діють під керівництвом визнаного лідера.

Також технологію було апробовано для побудови мережі співавторів з напрямку юриспруденції та напрямку – фізична оптика.

2. Мережу предметних областей було побудовано для наступних понять:
  - Physical optics
  - Scientometric.

Для мережі предметних областей за напрямком – фізична оптика:

- Кількість вузлів – 401;
- Кількість ребер – 670;
- Середня вага вузлу 1, 67;
- Діаметр мережі – 6;
- Середній найкоротший шлях – 2,48.

Частина побудованої мережі зображено на рисунку 8.

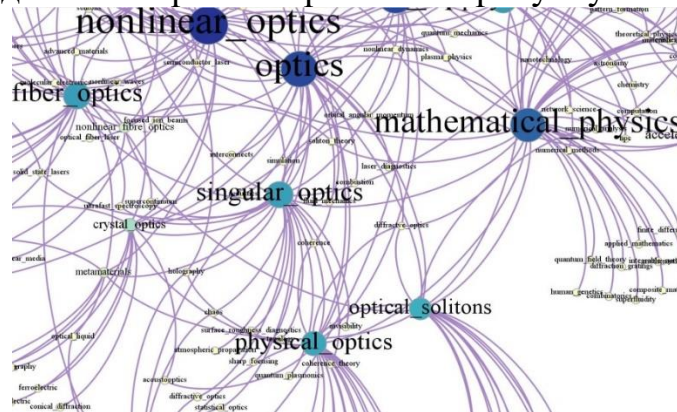


Рисунок 8 – Частина мережі предметних областей за напрямком – фізична оптика

3. Вікі-індекс автора на базі енциклопедичного ресурсу Wikipedia було побудовано для трьох науковців, і отримано відповідно наступні результати:
  - Альберт Енштейн:  $WI=12*11,5=141$ , (138 статей з бібліографією,  $wh=12$ );
  - Енрико Фермі:  $WI=7*9,6=67$ , (92 статті з бібліографією,  $wh=7$ );
  - Бенуа Мандельброт:  $WI=6*\sqrt{11}=19,9\sim 20$ , (11 статей з бібліографією,  $wh=6$ ).

Також було проведено порівняльний аналіз отриманого показника із притаманним в наукометрії для різних наукометричних систем індексами (таблиця 1).

Таблиця 1 – Аналіз показника індексу популярності для заданих науковців

| N п/п | Вчений            | Вікі-індекс | Індекс Гірша Scopus | Індекс Гірша Web of Science | Індекс Гірша Google Scholar |
|-------|-------------------|-------------|---------------------|-----------------------------|-----------------------------|
| 1.    | Albert Einstein   | 141         | 36                  | 6                           | 110                         |
| 2.    | Enrico Fermi      | 67          | 26                  | 1                           | 49*                         |
| 3.    | Benoit Mandelbrot | 20          | 31                  | 36                          | 90                          |

\*З огляду на відсутність профілю автора, індекс було обраховано для «e.fermi».

4. Побудову онтології на базі ресурсу Wikipedia було реалізовано для понять та для власного імені.

- Було побудовано онтологію для поняття ‘Sceintometric’, ‘mission statement’ та для імені ‘Evgen Paton’.

5. Мережу предметних областей на базі ресурсу препринтів arXiv було реалізовано для поняття ‘cavitation’ та ‘Ukraine’ і отримано наступні результати.

Задане для пошуку поняття ‘cavitation’ було відображено у 254 публікаціях з наступних наукових напрямків: фізика; комп’ютерні науки; кількісна біологія; математика; статистика.

Для притаманної предметної області було обраховано показник Term Frequency, що дорівнює – 5.

В дисертаційній роботі було апробовано методику наукометричного аналізу, що включає методи аналізу складних мереж, методи фільтрації тексту, візуалізацію даних за допомогою програмних засобів Gephi.

## ВИСНОВКИ

За результатами дисертаційної роботи було розв'язано ряд задач, що ставлять на меті оптимізувати роботу користувача із науковою інформацією – оптимізувати час роботи із великими обсягами даних і отримувати на виході набори та масиви даних, придатних для формування стратегії досліджень, прийняття управлінських рішень і пошуку міждисциплінарної ланки для пошуку точок дотику і різними предметними областями стосовно формування успішних колаборацій для реалізації складних мультидисциплінарних задач.

У рамках роботи відповідно до поставлених задач було отримано такі результати.

1. У роботі розроблено та реалізовано алгоритми для розширення можливостей отримання нових інформаційних масивів на базі ресурсу Google Scholar, а саме – побудова мережі предметних областей та мережі співавторів. Показано, що обсяг отриманих масивів інформації залежить від кількості інформації, що була надана користувачами ресурсу при оформленні власного профілю та міцності зв'язків між параметрами мережі. Саме такі результати є початковою ланкою при формування дослідницьких груп для реалізації масштабних проектів. Інструменти побудови мережі співавторів та мережі предметних областей на базі ресурсу Google Scholar є одним із підходів для формування експертних груп для оцінки наукових проектів, результатів реалізації

грантових програм, рецензії публікаційних доробків науковців, незалежних експертів для розв'язку суперечливих наукових питань. З огляду на те, що ресурс є динамічним і щоденно розширюється та оновлюється за рахунок додавання нової інформації – нових публікацій і співставлення її із авторами – зареєстрованими користувачами системи і водночас актуалізації інформації користувачами щодо місця проведення досліджень, нових співавторів, нових наукових напрямків реалізації наукових ідей.

2. Для визначення видимості та значущості для широкого кола користувачів Інтернет розроблено індекс популярності автора (науковця) на базі енциклопедичного ресурсу відкритого доступу Вікіпедія. Розроблений індекс було опробовано для провідних світових науковців різних галузей знань і також співставлено із найбільш уживаними на сьогодні індексами різних наукометричних ресурсів. Використання даного індексу дозволить оцінити видимість наукових персоналій того чи іншого закладу для суспільства і відповідно – популяризації наукових досягнень та пов'язаних з ними постатей.

3. Розроблено та реалізовано алгоритм побудови онтології на базі енциклопедичного ресурсу відкритого доступу Вікіпедія. Алгоритм було апробовано на поняттях, що позначають як терміни, так і власні імена. Показано, що за рахунок реалізації алгоритму можна отримати дані, що можуть продемонструвати широту понять, які можуть бути пов'язані між собою. У той же час такий підхід може збільшити коло статей в он-лайн енциклопедії, в яких може бути присутня інформація, та відповідні посилання, як про організацію, науковців, наукових груп та досліджень відповідно до взаємозв'язків базових понять стосовно досліджень, що провадяться.

4. За рахунок розробки моделі «Концепт – масив наукових публікацій» та побудова відповідного алгоритму було забезпечено повноту формування мережі предметних областей для поняття. Реалізація алгоритму показала широкий спектр застосування заданого поняття в різних предметних областях. Реалізовані моделі та програмна реалізація алгоритмів дозволить здійснювати пошук міждисциплінарної складової для пошуку партнерів та формування успішних колаборацій для реалізації складних міждисциплінарних проєктів для вирішення задач, які можуть бути вирішені лише за рахунок синергії зусиль представників різних наукових напрямків та наукових шкіл.

Дослідження, представлені в роботі, можуть бути розвинені шляхом розробки та формування універсальних моделей, що можуть бути застосовані до різних ресурсів наукової та наукометричної інформації. Програмна реалізація відповідних алгоритмів може бути здійснена різними мовами програмування, зокрема тими, що орієнтовані на роботу із великими масивами даних. Також можливе представлення он-лайн додатків для доступу широкого кола користувачів до можливостей представлених в рамках роботи.

Обґрунтовано можливість застосування розроблених алгоритмів як додаткових інструментів наукометричного аналізу. Показано важливість інформаційних масивів, що формуються на базі ресурсів, які містять бібліометричну та наукометричну інформацію і можуть виступати параметрами,

що можуть впливати на прийняття рішення та розширення спектру можливої співпраці між науковцями.

## СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Ланде ДВ, **Андрущенко ВБ**, Балагура ІВ. Вікі-індекс популярності авторів наукових публікацій. Реєстрація, зберігання і обробка даних. 2016. 18 (4): с. 44-54.
2. Ланде ДВ, **Андрущенко ВБ**. Побудова мережі предметних областей на базі ресурсу архів. Реєстрація, зберігання і обробка даних. 2018. 20(2): 12-22.
3. **Андрущенко ВБ**. Нові інформаційні технології пошуку і обробки даних ресурсу препринтів архів. Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія «Технічні науки». 2018. 29(68), 3:84-89.
4. **Андрущенко ВБ**. Підходи до визначення критеріїв для аналізу on-line ресурсів наукової інформації. Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія «Технічні науки». 2018. 29(68), 4:84-89.
5. Гриньов БВ, Кияк БР, **Андрущенко ВБ**. Моніторинг активності вітчизняних учених через призму конкурсної діяльності. Вісник Національної академії наук України. Березень 2017. 3/2017: 75-81.
6. **Андрущенко ВБ**, Кияк БР. Обґрунтування критеріїв оцінювання фундаментальних наукових досліджень. Наука та наукознавство. Грудень 2015; 4(89):67-72.
7. Ланде ДВ, **Андрущенко ВБ**. Побудова мереж співавторства фахівців з юриспруденції за даними сервісу Google Scholar Citations. Інформація і право. 3/2016. 1(16): 146-150.
8. Красовська ОВ, **Андрущенко ВБ**, Величко ІГ. Освіта й наука та їхня роль у соціальному та індустріальному розвитку суспільства. Київ: Логос; 2015. Україно-німецьке наукове співробітництво в галузі фундаментальних досліджень (досвід Державного фонду фундаментальних досліджень України); с. 74-81.
9. **Андрущенко ВБ**, Кияк БР. Анотований збірник проектів спільного конкурсу ДФФД - БРФФД. Київ: ВД «Академперіодика»; 2017. Частина 1, Критерії та показники успішної міжнародної наукової співпраці; с. 5-9.
10. **Андрущенко ВБ**. Інформаційно-аналітична діяльність Державного фонду фундаментальних досліджень - важливий елемент формування національного наукового простору. В: Попик ВІ. Матеріали міжнародної науково-практичної конференції Місце і роль бібліотек у формуванні національного інформаційного простору. Національна бібліотека України ім. В.І. Вернадського; НБУВ; 2014, с. 208-210.
11. Ланде ДВ, **Андрущенко ВБ**, Балагура ІВ. Построение сетей соавторства по данным сервиса Google Scholar Citations. В: Голенков ВВ. Материалы

Міжнародної конференції Open Semantic Technologies for Intelligent Systems; 18-20 лютого 2016 року. Білоруський державний інститут інформатики та радіоелектроніки; БДУІР; 2016, с. 233-238.

12. **Андрущенко ВБ**, Ланде ДВ. Побудова онтології за допомогою сканування ресурсів Wikipedia. В: Литвиненко ОС. Матеріали VIII Міжнародної науково-технічної конференції «Інтелектуальні технології лінгвістичного аналізу»; 25-26 жовтня 2016 року; Національний авіаційний університет. Київ. Київ: НАУ; 2016, с. 10.
13. Балагура ІВ, **Андрущенко ВБ**. Аналіз інноваційних напрямів у педагогіці з огляду на публікаційну активність українських науковців. В: Вакаренко ОГ, редактор. Матеріали конференції «Наука України у світовому інформаційному просторі»; 2016; Київ. Академперіодика; 2016, с. 95-102.
14. **Андрущенко ВБ**. Порівняльний аналіз структур і реалізації пошуку наукометричних ресурсів з метою складання унікальних алгоритмів розширення можливостей існуючих систем. В: Інститут проблем реєстрації інформації НАН України. Матеріали конференції Реєстрація, зберігання та обробка інформації; Травень 2016; Київ. ІПІ, 2016, с. 110-111.
15. **Andrushchenko VB**, Lande DV. Sounding of Google Scholar Citations service as a way to obtain new scientometric data. В: Писаренко АВ. Summer InfoCom Advanced Solutions 2016: Date; Київ. Видавництво; 2016, с. 66-68.
16. Ланде ДВ, **Андрущенко ВБ**. Нові наукометричні сервіси на базі Google Scholar Citations. В: Панкратова НД. System Analysis and Information Technologies 18-th International Conference SAIT 2016; Institute for Applied System Analysis NTUU “KPI”, 2016, с. 52.
17. **Андрущенко ВБ**, Балагура ІВ, Ланде ДВ. Інформаційні ресурси доступу та обміну науковою інформацією, системи ідентифікації науковців - можливості, недоліки, переваги. В: Додонов АГ. Матеріали міжнародної науково-технічної конференції Інформаційні технології та безпека. 2 грудня 2016 року. ІПІ; 2017, с. 180-191.  
**Andrushchenko VB**, Balagura IV, Lande DV. Information Resources for Scientific Information Access and Exchange, Identification of Scientists – Opportunities, Disadvantages, Benefits. In: CEUR Workshop Proceedings. Kyiv, Ukraine. 2016 Dec 1. Vol.1813: 62-67.
18. Lande DV, **Andrushchenko VB**, Balagura IV. Formation of the Subject Area on the Base of Wikipedia Service. В: Голенков ВВ. Матеріали Міжнародної конференції Open Semantic Technologies for Intelligent Systems; 17-19 лютого 2017 року. Білоруський державний інститут інформатики та радіоелектроніки; БДУІР; 2017, с. 211-215
19. Lande DV, **Andrushchenko VB**, Balagura IV. An Index of Authors’ Popularity for Internet Encyclopedia. В: Національний лісотехнічний університет України. The 1st International Conference COMPUTATIONAL LINGUISTICS AND INTELLIGENT SYSTEMS (COLINS 2017); Дата. 21 квітня 2017 року. Видавець Національний технічний університет «Харківський політехнічний інститут». 2017, с. 47-55.

20. **Андрущенко ВБ**. Побудова дерева предметних областей для заданого поняття на базі ресурсу препринтів ArXiv. В: Литвиненко ОЄ. Матеріали XI Міжнародної науково-технічної конференції «Інтелектуальні технології лінгвістичного аналізу»; Дата 24-25 жовтня 2017 року; Національний авіаційний університет. Київ. Київ: НАУ; 2017, с. 20.
21. **Андрущенко ВБ**, Балагура ІВ. Аналіз публікаційної активності за напрямком комп'ютерної безпеки на базі ресурсів Web of Science та Scopus. В: Додонов ВГ. Матеріали міжнародної науково-технічної конференції Інформаційні технології та безпека. 29-30 листопада 2017 року. ПІПІ; 2017, с. 8-17.
- Andrushchenko VB**, Balagura IV. Analysis of Publication Activity in the Area of Computer Security Based on Web of Science and Scopus Data. In: CEUR Workshop Proceedings. Kyiv, Ukraine. 2017 Nov 30. Vol.2067: 8-15.
22. Ланде ДВ, **Андрущенко ВБ**, Wikipedia Index of Scientist's Popularity. В: Дичка ІА. XVII Міжнародна наукова конференція імені Т.А. Таран "Інтелектуальний аналіз інформації. IAI2017, Київ, 17-19 травня 2017 р. Просвіта, 2017. с. 137-143.
23. Lande D, **Andrushchenko V**, Balagura I. Data Science in Open-Access Research On-line Resources. Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining and Processing; 2018 August 21-25; Lviv, Ukraine. p. 17-20.
24. Lande DV, **Andrushchenko VB**. Formation of subject area and the co-authors network by sounding of Google Scholar Citations service. Arxiv.org. arXiv:1605.02215. 2016. Available from: <https://arxiv.org/abs/1605.02215>
25. Lande DV, **Andrushchenko VB**, Balagura IV. Wiki-index of authors popularity. Arxiv.org. 1702.04614. 2017. Available from: <https://arxiv.org/abs/1702.04614>

## АНОТАЦІЯ

**Андрущенко В.Б. Інформаційні технології наукометричного аналізу на основі моніторингу ресурсів мережі Інтернет. – Рукопис.**

Дисертація на здобуття наукового ступеня кандидата наук за спеціальністю 05.13.06 – Інформаційні технології. – Тернопільський національний технічний університет імені Івана Пулюя, Тернопіль, 2019.

Дисертаційна робота присвячена огляду та аналізу ресурсів наукової інформації, дослідженню можливостей розробки додаткових інформаційних масивів на базі існуючих джерел наукової та наукометричної інформації. В рамках роботи запропоновано підходи до визначення параметрів для проведення оцінки ресурсів наукової та наукометричної інформації, способи отримання нових даних, що дозволить оптимізувати час роботи і даними та залучати менше ресурсів для реалізації робіт із пошуку наукової та наукометричної інформації для подальшої її обробки.

В роботі представлено та реалізовано алгоритми для побудови мережі співавторів та предметних областей на базі ресурсу Google Scholar Citations.

Показано способи апробації алгоритмів із відповідною візуалізацією. На базі енциклопедичного ресурсу Вікіпедія запропоновано новий індекс для визначення видимості та цитованості автора (науковця), а також алгоритм побудови онтології за зазначеним поняттям. Показано можливість реалізації технології для поняття – визначення, а також власного імені. Представлено можливості отримання інформації за рахунок опрацювання результатів пошуку за заданим поняттям в рамках ресурсу препринтів arXiv.

Отримані результати окреслюють повний спектр притаманності заданого для пошуку поняття предметним областям і також визначено перелік галузей, для яких поняття є найбільш уживаним. Показано важливість інформаційних масивів, що формуються на базі ресурсів, які містять бібліометричну та наукометричну інформацію і можуть виступати параметрами, що можуть впливати на прийняття рішення та розширення спектру можливої співпраці науковців, формування унікальних наукових колаборацій.

**Ключові слова:** мережі співавторів, онтології, індекс популярності науковця, мережа предметних областей, Google Scholar Citations, Вікіпедія, arXiv, наукометричні показники

## АННОТАЦИЯ

**Андрущенко В.Б. Информационные технологии наукометрического анализа на основе мониторинга ресурсов сети Интернет. – Рукопис.**

Диссертация на соискание научной степени кандидата наук по специальности 05.13.06 – Информационные технологии. – Тернопольский национальный технический университет имени Ивана Пулюя, Тернополь, 2019.

Диссертационная работа посвящена обзору и анализу ресурсов научной информации, исследованию возможностей разработки дополнительных информационных массивов на базе существующих источников научной и наукометрической информации. В рамках работы предложены подходы к определению параметров для проведения оценки ресурсов научной и наукометрической информации, способы получения новых данных, что позволит оптимизировать время работы с информацией и привлекать меньше ресурсов для реализации работ по поиску научной и наукометрической информации и последующей ее обработки.

В работе представлены и реализованы алгоритмы для построения сети соавторов и предметных областей на базе ресурса Google Scholar Citations. Показаны способы апробации алгоритмов с соответствующей визуализацией. На базе энциклопедического ресурса Википедия предложен новый индекс для определения видимости и цитируемости автора (ученого), а также алгоритм построения онтологии для указанного понятия. Показана возможность реализации технологии для понятия - определение, а также имени собственного. Представлены

возможности получения информации за счет обработки результатов поиска по заданным понятием в рамках ресурса препринтов arXiv.

Полученные результаты определяют полный спектр присущих заданного для поиска понятия предметным областям и также определен перечень отраслей, для которых понятие является наиболее употребляемым. Показана важность информационных массивов, которые формируются на базе ресурсов и содержат библиометрическую и наукометрическую информацию и могут выступать параметрами, которые способны влиять на принятие решения и расширения спектра возможного сотрудничества ученых, формирование уникальных научных коллабораций.

**Ключевые слова:** сети соавторов, онтологии, индекс популярности ученого, сеть предметных областей, Google Scholar Citations, Википедия, arXiv, наукометрические показатели.

## ABSTRACT

**Andrushchenko V.B. Information technologies of scientometric analysis based on monitoring of the Internet recourses.**

Thesis for Ph.D. degree in technical science on specialty of 05.13.06 – Information technologies. – Ternopil Ivan Pul`uj National Technical University, Ternopil, 2019.

The thesis is devoted to the review and analysis of scientific information resources, the study of the possibilities of developing additional information arrays based on existing sources of scientific and scientometric information. As part of the work, approaches to determine parameters for assessing the resources of scientific and scientometric information, methods of obtaining new data sets are proposed, which will optimize the time of working with the information and attract less resources to implement work on finding scientific and scientometric information and its subsequent processing.

The paper presents and implements algorithms for building a network of co-authors and subject areas based on the Google Scholar Citations resource. The ways of approving algorithms with appropriate visualization are also represented. Based on the open-access encyclopedic resource Wikipedia, a new index has been described for determining the visibility and quoting of the author (researcher), as well as an algorithm for constructing of ontology according to the indicated concept. The possibility of implementing the technology for the concept - definition, as well as the proper name is described. The possibilities of obtaining information through the processing of search results for a given concept within the framework of the ArXiv preprint resource are presented.

The obtained results determine the full range of inherent in the concept defined for the search for subject areas and also the list of industries for which the concept is the most used is defined. The importance of information arrays, which are formed on the basis of resources and contain bibliometric and scientometric information, and can act as

parameters that are able to influence the decision making and expand the range of possible cooperation of researchers, the formation of unique scientific collaborations.

**Key words:** co-authors networks, ontology, index of researcher' popularity, subject domain network, Google Scholar Citations, Wikipedia, arXiv, scientometric indexes.