Інститут проблем реєстрації інформації

Національна академія наук України

Тернопільській національний технічний університет імені Івана Пулюя

Міністерство освіти і науки України

Кваліфікаційна наукова праця

на правах рукопису

АНДРУЩЕНКО Валентина Борисівна

УДК 004.031.42

ДИСЕРТАЦІЯ

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ НАУКОМЕТРИЧНОГО АНАЛІЗУ НА ОСНОВІ МОНІТОРИНГУ РЕСУРСІВ МЕРЕЖІ ІНТЕРНЕТ

05.13.06 – Інформаційні технології
Технічні науки

Подається на здобуття наукового ступеня кандидата технічних наук

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело
_____ Андрущенко В.Б.

Науковий керівник: ЛАНДЕ Дмитро Володимирович
доктор технічних наук, старший науковий співробітник

Київ – 2018

АНОТАЦІЯ

Андрущенко В.Б. Інформаційні технології наукометричного аналізу на основі моніторингу ресурсів мережі інтернет – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 «Інформаційні технології». – Тернопільський національний технічний університет імені Івана Пулюя, Тернопіль, 2018.

Робота виконана в Інституті проблем реєстрації інформації НАН України. Захист відбудеться на засіданні спеціалізованої вченої ради К58.052.06 в Тернопільському національному технічному університеті імені Івана Пулюя.

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 підходти до визначення параметрів для провдення оцінки ресурсів наукової та наукометричної інформації, способи отримання нових масивів інформації, що дозволить оптимізувати час роботи із данними та залучати менше ресурсів для реалізації робіт із пошуку наукової та наукометричної інформації і подальшої її обробки.

Сучасне наукове суспільство висуває нові виклики, що мають на меті залучити якомога більше індикаторів для оцінки, систематизації та узагальнення наукової роботи. Одним із підходів, що загальноприйняті у світі і застосовуються для формування загальної картини визначеної галузі та науки країни в цілому, оцінки публікаційної активності, успішності діяльності вченого, наукового колективу, проекту, окремої інституції є застосування наукометричних показників. Для організації роботи із наукометричним інструментарієм розроблено ряд ресурсів, що представляють таку інформацію: переліки публікацій, видань, галузева приналежність, відповідні показники, що обраховуються автоматично.

Питання аналізу ресурсів, або визначення критеріїв для аналізу ресурсів зосереджено на особливостях розробки критеріїв для оцінювання електронних освітніх ресурсів з метою підвищення якості знань, що набуває користувач.

З огляду на актуальність поставленої задачі та наявності в глобальних мережах великої кількості ресурсів наукової інформації, важливим є питання систематизації ресурсів та визначення основних характеристик з огляду на наповнення та спосіб розробки систем – достатність інформації та врахування орієнтованості щодо користувача.

Визначення критеріїв для аналізу та оцінки ресурсів наукової та наукометричної інформації надає можливість також здійснити опис ресурсів наукової інформації з точки зору її доступності та наповнення. Це в свою чергу дозволить не тільки орієнтуватися у масиві ресурсів, що наразі є доступними користувачам мережі Інтернет, а й вдосконалювати існуючі системи, з огляду на вимоги щодо розробки інформаційних систем та користувацького інтерфейсу зокрема, та визначати перелік даних відповідно до задачі тієї чи іншої системи.

Задачами дисертаційного досілдження також є розширення можливостей існуючих систем наукової та наукометричної інформації для формування нових масивів наукової інформації для подальшого їх використання не тільки як допоміжного інструментарію, а й джерела необхідної інформації для формування стратегії наукового досілдження та розвитку.

В рамках роботи було запропоновано моделі для побудови мережі співавторів та предметних областей на базі ресурсу Google Scholar Citations.

В запропонованій моделі для побудови мережі співавторів використовуються лише базові теги. Надалі алгоритм передбачає використання знань, що закладені співавторами, теги позначені як головні для них – експертне середовище в цьому випадку істотно розширюється. Необхідно зазначити, що система Google Scholar Citations є зручною щодо доступу до інформації, не передбачає створення власного профілю користувача для доступу до інформації, доступ є необмежений. В той же час кожен зареєстрований має можливості збереження отриманої і скорегованої інформації.

Показано, що переваги запропонованих моделей полягають в тому, що участь експерта є мінімальною – зазначення базових термінів для реалізації поставленої задачі. За рахунок апробації алгоритмів та програмних додатків було показано обсяг результатів, що можуть бути отримані за рахунок сканування ресурсів наукової та наукометричної інформації. Крім того було представлено візуалізацію отриманих результатів.

В роботі пртділено увагу розробці підходів до отримання нових даних найбільш популярного в мережі інтернет енциклопедичного ресурсу Вікіпедія.

Енциклопедичний ресурс Вікіпедія було обрано для розгляду з огляду на його доступність користувачам мережі інтренет – зокрема можливість формування надбудов. Ресурс є абсолютно відкритий для користувачів, не потребує передплати та додаткової реєстрації користувача для доступу до інформації. Статті, що містить ресурс формуються та редагуються авторами-дописувачами-користувачами із зазначенням посилань на відповідні джерела, що використовуються для формування текстових масивів із різноманітної тематики.

В роботі запропоновано новий індекс для визначення видимості та цитованості автора (науковця) у енциклопедичному ресурсі відкритого доступу Вікіпедія. На відміну від існуючих наукометричних показників, що з одного боку враховують як кількісну – кількість пубілкацій і кількість цитувань, так і якісну – які саме видання цитуються, запропонований індекс запропоновано обраховувати за рахунок даних, що представлені відкритою енциклопедією і сформовані за рахунок використання, як наукових джерел, так і науково-популярної літератури. Саме це – підкреслює значущу складову, що демонструє вплив результатів досліджень на суспільство – значною частиною якого є саме користувачі ресурсів відкритого доступу – зокрема Вікіпедія.

Також здобувачем було запропоновано новий підхід до побудови онтології за визначеним поняттям на базі енциклопедичного ресурсу Вікіпедія. Показано можливість реалізації технології для поняття – визначення, поняття, що представлене словом, словосполученням, а також власним іменем. Від статичних моделей предметних областей такий підхід відрізняється урахуванням динамічної

зміни контенту бази даних цього сервісу, урахуванням нових понять та нової інформації, що додається до існуючих статей за рахунок редагування та доповнення інформації різними авторами - користувачами Вікіпедія.

Реалізовані задачі – практичне застосування отриманих результатів було представлено в роботі. Також результати візуалізовано у вигляді мереж в середовищі Gephi.

З огляду на те, що сучасне суспільство висуває все нові вимоги не тільки до якості робіт, а й до часу, з яким вони стають доступні науковому загалу – актуальним на сьогодні є розміщення науковий текстів, що підготовлені і направлені на розгляд до наукових видань, у репозитаріях та архівах препринтів. Саме з огляду на час публікації наукових статей в друкованих виданнях чи в мережі інтернет зробило авторитетні (наявність індексу DOI або узагальненої попередньої рецензії) ресурси препринтів та репозитарії джерелом публікаційного доробку, на який можна посилатися при поданні запитів на гранти, зокрема Рамкової програми з досліджень та інновацій Європейського Союзу «Горизонт 2020» – гранти Європейської ради досліджень.

Орієнтований, при стовренні, тільки на роботи із фізики, архів препринтів arXiv регулярно розширюються за рахунок додавання нових наукових напрямків, за якими можна розмістити свої публікаційні доробки на ресурсі.

Здобувачем в роботі запропоновано проаналізувати масив публікацій, що міститься в архіві з точки зору приналежності їх до наукових напрмків із використанням лише інформації самого ресурсу – предметна область – наковий напрямок, і задавати для пошуку лише певний концепт.

Перевагою поряд із іншими інформаційними ресурсами, що були розглянути в роботі – відкритість архіву та відповідні ліцензії дозволяють відкрито користуватися даними і формувати програмні надбудови для роботи з системою і отримання нової інформації та аналітичних масивів на базі ресурсу.

Із використанням методів математичної лінгвістики формуються словники, що містять інформацію про предметні області і відповідні наукові напрямки передбачені системою.

Отримані результати окреслюють повний спектр притаманності заданого для пошуку поняття предметним областям і також визначено перелік галузей, для яких поняття є найбільш уживаним.

Запропоновані підходи до формування нових інформаційних масивів та отримані результати є актуальним інструментарієм для роботи як користувачів ресурсами науково-технічної та наукометричної інформації, так і користувачів мережі Інтернет. З одного боку – робота із енциклопедичними ресурсами відкритого доступу розрахована на всіх користувачів Інтернет і результати роботи будуть корисні як, науковцям, так і носити популяризаційне навантаження.

Дані, що отримані за результатами роботи – є інструментом для науковця, адміністратора науки, керівника установи, представників центральних органів влади при формуванні політики роботи, стратегії розвитку досілджень, способи створення колаборацій для реалзації наукових ідей в рамках міждсциплінарної складової великих проектів.

Показано важливість інформаційних масивів, що формуються на базі інформаційних ресурсів, які містять бібліометричну та наукометричну інформацію і можуть виступати тими параметрами, що можуть впливати на прийняття рішення та розширення спектру можливої співпраці науковців, формування унікальних науково-дослідних колаборацій.

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

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, 2018.

The work was provided in the Institute for Information Recording of the National Academy of Sciences of Ukraine. The defense of the dissertation will take place in frames of the Academic Council meeting K58.052.06 in Ternopil Ivan Pul`uj National Technical University.

The work is dedicated to the overview and analysis of the research information resources, studying of the opportunities to develop additional information arrays based on the existing sources of the research and scientometric information.

In frames of the work there offered attempts to defining of the parameters for providing the assessment of the research and scientometric information, the ways of obtaining of new information arrays which will allow optimizing of the time of the data procession and involving less resources for the research and scientometric search procedures and its further processing.

Modern scientific society creates new challenges that aim to involve as many as possible indicators for evaluation, systematization and consolidation of the research work. One of the world conventional approaches, which are applied for the forming of the general view of the defined domain and the science in general, evaluation of the publication activity, the impact of activity of researcher, research team, project, and separate institution, is the application of the scientometric indexes.

For organization of the work with the scientometric instruments there developed the range of instruments that can represent the next information: the list of publications, journals, particular branch belonging, appropriate parameters that are calculated automatically.

For today, the issue of the resources analysis or the definition of the criteria for the resources analysis is mostly concentrated on the peculiarities of the figuring out of the

criteria for evaluating of the educational e-resources in order to increase the quality knowledge that the user obtains.

From the point of view of the objective and the availability of the quantity of the research resources in the Internet, the core issue is the systematization of resources and definition of the main characteristics form the point of view of the populating and the ways of the systems development – the information sufficiency and considering the user.

Identification of criteria for analysis and evaluation of the research and scientometric information resources gives an opportunity to provide the description of the research information resources form the point of view of their accessibility and content. It will allow not only to navigate in the range of available for the Internet users resources, but also to improve the existing systems from the point of view of the development of the information systems and users interface in particular, and to determine the list of data according to the objective of the system.

The tasks of the dissertation are also to expand the capabilities of existing systems of scientific and scientific information to form new arrays of scientific information for their further use, not only as auxiliary tools, but also sources of necessary information for the formation of a strategy for scientific development and growth. Within the the work, models were proposed for building a network of co-authors and subject areas based on the Google Scholar Citations resource.

In the proposed model, only basic tags are used for constructing a co-authoring network. The algorithm involves the use of knowledge, laid by co-authors, tags are designated as the main ones for them - the expert environment in this case significantly expands. It should be noted that the Google Scholar Citations system is convenient for access to information, it does not require the creation of a user profile for access to information, access is unlimited. At the same time, each registered has the ability to save received and corrected information.

It is shown that the advantages of the proposed models are that the participation of the expert is minimal - the indication of the basic terms for the realization of the task.

Due to approbation of algorithms and software applications, the scope of results that can be obtained by scanning resources of scientific and scientometric information is shown.

In addition, visualization of the results was presented. The paper focuses on the development of approaches to obtaining new data on the most popular online encyclopedia resource Wikipedia.

The Wikipedia encyclopedia resource was chosen for consideration because of its availability to Internet users, in particular the ability to create add-ons. The resource is completely open to users, does not require subscription and additional user registration for access to information. Articles are generated and edited by the authors-contributors-users, with references to the relevant sources used to form text arrays on a variety of topics.

The paper proposes a new index for determining the visibility and citation of the author (scientist) in the encyclopedia of open access Wikipedia. In contrast to the existing scientometric indicators, which, on the one hand, take into account both quantitative - the number of issues and the number of quotes, as well as the qualitative ones - which publications are cited, the proposed index is aimed to be calculated from the data presented by the open encyclopedia and formed through the use of as scientific sources, and popular science literature.

This is what emphasizes a significant component that demonstrates the impact of research results on society - a large part of which is precisely the users of open access resources - in particular Wikipedia.

Also, the applicant proposed a new approach to the construction of ontology for a definite concept on the basis of the encyclopedia resource Wikipedia. It is shown the possibility of implementing technology for the concept - the definition, the concept represented by the word, the phrase, as well as its own name. From static models of subject areas, this approach differs by taking into account the dynamic change of the content of the database of this service, taking into account new concepts and new information, which is added to existing articles by editing and supplementing information by various authors - Wikipedia users.

Realized tasks - practical application of the obtained results was presented in the work. Also, the results are rendered as networks in the Gephi environment.

Taking into account the fact that modern society puts forward more demands not only on the quality of work, but also on the time with which they become available to the scientific community, today the publication of scientific texts prepared and directed for consideration to scientific publications is relevant today. repositories and archives of preprints.

It is precisely in view of the time of the publication of scientific articles in the print media or on the Internet that authoritative (the existence of the DOI index or generalized prior review) made the resources of the preprints and repositories a source of publication, which can be referenced when submitting grant applications, in particular the Framework Program for Research and Innovations of the European Union "Horizon 2020" - grants from the European Research Council. Oriented, with staging, for work on physics only, the archives of preprints are regularly expanded by adding new scientific directions that can accommodate their publications on the resource.

The applicant in the work proposed to analyze the array of publications contained in the archive in terms of their belonging to scientific sources using only the information of the resource itself - the subject area - the scientific direction, and ask for the search only a certain concept. The advantage, along with other information resources that were considered in the work - the openness of the archive and the corresponding licenses allow openly use the data and form a software add-in to work with the system and obtain new information and analytical arrays based on the resource.

Using the methods of mathematical linguistics dictionaries containing information on subject areas and corresponding scientific directions provided by the system are formed. The obtained results outline the full spectrum of the peculiarity of the given object for the search for the concept and the list of industries for which the concept is most used is defined.

The offered approaches to the formation of new information arrays and the results obtained are an actual tool for working as users of the resources of scientific, technical,

scientific and mathematical information, as well as users of the Internet. On the one hand, work with encyclopedic open access resources is intended for all users of the Internet and the results of the work will be useful as for scholars, as well as carry a popularization objective.

The data obtained by the results of the work - is an instrument for a scientist, administrator of science, head of the institution, representatives of central authorities in shaping the policy of work, development strategies, ways of creating collaborations for the implementation of scientific ideas within the framework of the interdisciplinary component of large projects.

It is shown the importance of information arrays that are formed based on information resources that contain bibliometric and scientometric information and can serve as parameters that can influence the decision-making process and expand the range of possible cooperation of scientists, the formation of unique research collaborations.

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

Список публікацій здобувача

1. Ланде ДВ, Андрущенко ВБ, Балагура ІВ. Вікі-індекс популярності авторів наукових публікацій. Реєстрація, зберігання і обробка даних. 2016. 18 (4): с. 44-54. (Особистий внесок – оцінка можливостей використання бібліографічних посилань статей Wikipedia для виокремлення інформації про авторів).

2. Ланде ДВ, Андрущенко ВБ. Побудова мережі предметних областей на базі ресурсу arxiv. Реєстрація, зберігання і обробка даних. 2018. 20(2): 12-22. (Особистий внесок – розробка алгоритму пошуку наукових публікацій за заданим концептом і формування мережі предметних областей на базі отриманих результатів, оцінка отриманих результатів)
3. Андрущенко ВБ. Нові інформаційні технології пошуку і обробки даних ресурсу препринтів arxiv. Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія «Технчні науки». 2018. 29(68), 3:84-89. (Особистий внесок – розробка моделі «Концепт – система наукових напрямків, розробка алгоритму для реалізації побудови мережі предметних областей, апробація результатів на заданому концепті)
4. Андрущенко ВБ. Підходи до визначення критеріїв для аналізу on-line ресурсів наукової інформації. Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія «Технчні науки». 2018. 29(68), 4:88-95. (Особистий внесок – розробка підходів до формування критеріїв для оцінки наукової інформації)
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. (Особистий внесок – визначення підходів до оцінки активності вітчизняних вчених за рахунок сканування великих масивів інформації про грантові пропозиції)
6. Андрущенко ВБ, Кияк БР. Обґрунтування критеріїв оцінювання фундаментальних наукових досліджень. Наука та наукознавство. Грудень

2015; 4(89):67-72. (Особистий внесок – розробка критеріїв із запропонованого переліку щодо оцінювання фундаментальних досліджень)

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. (Особистий внесок – участь у розробці алгоритму реалізації поставленої задачі)
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 фундамнтальних досліджень (досвід Державного фонду фундаментальних досілджень України); с. 74-81. (Особистий внесок – моніторинг масиву даних про проекти спільних конкурсів та виокремлення наукометричної інформації із провденням аналізу отриманих результатів)
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. (Особистий внесок – аналіз масиву даних про конкурсні проекти за заданим напрямком, викоремлення наукометричних даних результатів пошуку та їх аналіз)
10. Андрущенко ВБ. Інформаційно-аналітична діяльність Державного фонду фундментальних 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. (Особистий внесок – опис процесу розробки, вдосконалення та порядку роботи інформаційно-аналітичної системи ДФФД, підходи до аналізу даних, що містить система)
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. (Особистий внесок – участь у розробці алгоритму для реалізації поставленої задачі)

12. Андрущенко ВБ, Ланде ДВ. Побудова онтології за допомогою сканування ресурсів Wikipedia. В: Литвиненко ОЄ. Матеріали VIII Міжнародної науково-технічної конференції «Інтелектуальні технології лінгвістичного аналізу»; 25-26 жовтня 2016 року; Національний авіаційний університет. Київ. Київ: НАУ; 2016, с. 10. (Особистий внесок – розробка та програмна реалізація алгоритму для вирішення задачі)
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. (Особистий внесок – виокремлення та аналіз результатів за заданою тематикою розробленими програмними засобами)
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. ІПРІ, 2016, с. 110-111. (Особистий внесок – розробка підходів до порівнння інформаційних систем наукометричної іфнормації)
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. (Особистий внесок – оцінка масиву інформації, що містить система та розробка алгоритмів для виокремлення інформації для формування нових масивів даних)
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, c. 52. (Особистий внесок – оцінка масиву інформації, що містить система та розробка алгоритмів для виокремлення інформації для формування нових масивів даних)

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. ІПРІ; 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.

    (Особистий внесок – визначення параметрів для проведення моніторингу, аналізу та узагальнення інформації про ресурси мережі Інтернет, що містять наукову інформацію)

18. Lande DV, AndrushchenkoVB, Balagure IV. Formation of the Subject Area on the Base of Wikipedia Service. В: Голенков ВВ. Материалы Міжнародної конференції Open Semantic Technologies for Intelligent Systems; 17-19 лютого 2017 року. Білоруський державний інститут інформатики та радіоелектроніки; БДУІР;2017, с. 211-215. (Особистий внесок – розробка алгоритму та його програмна реалізація)

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. (Особистий внесок – участь у розробці алгоритму для реалізації поставленої задачі)

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. (Особистий внесок – розробка моделі, алгоритму, програмна реалізація та апробація результатів дослідження)
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. ІПРІ; 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.
(Особистий внесок – моніторинг системи наукометричної інформації для виокремлення масиву публікацій за заданим напрямком та подальший аналіз розробленими програмними засобами)
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. (Особистий внесок – участь у розробці алгоритму для реалізації поставленої задачі)
23. Lande D, AndrushchenkoV, 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. (Особистий внесок – розробка алгоритму та апробація результатів дослідження)
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 (Особистий внесок – участь у розробці алгоритму для реалізації поставленої задачі)

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. (Особистий внесок – участь у розробці алгоритму для реалізації поставленої задачі)

ЗМІСТ

ВСТУП

Обгрунтування вибору теми дослідження. На сьогодні наукометрична інформація все більше стає інструментом для застосування інформаційних технологій, наприклад, для оцінки наукової впливовсті, 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 інформаціними 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 передплачений, 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 разноманітних 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 зовнушньої оцінки наукових рішень, проектів, 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. Біскаро, 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 взяємозв’язок між науковими напрямками та предметними областями з огляду на актуальність міждисципдінарної 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 можливіть 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 конркетного 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 популяризції 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 викоремлення необхідної інформації, отримати інструменти для прийняття рішень щодо визначення експертих груп для оцінки наукових проектів, результатів реалізації наукових досліджень, рецензування

наукових публікацій; отримати новий актуальний інструмент для пошуку партнерів для реалізації великих наукових проектів в рамках дослідницьких колаборацій поряд із звичними підходами і способи формуваня стратегії популяризації досліджень і доведення їх до широкого загалу споживачів наукової інформації.

Зв'язок роботи з науковими програмами, планами, темами, грантами. Результати дисераційної роботи та запропоновані у роботі методи, моделі та відповідні алгоритми відображено у науковому звіті Інституту проблем реєстрації інформації НАН України, відповідно до плана 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 проекту 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 обширу 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 лінвістики.

Методи дослідження. Для реалізації поставлених задач було використано такі наукові методи: методи аналізу текстових масивів, методи статистичного аналізу, методи математичної лінвістики, 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 реалізовних 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 елементам 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;
– розробку 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 викоганих у співавторстві, здобувачеві належить: [1, 2] – розробка моделей та нових індексів; [5, 6, 9] – розробка критеріїв для оцінки результативності фундаментальних досліджень, що грунтуються 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 проектів; [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 2енциклопедичний 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 предетних областей на базі ресурсу arxiv.

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-технійчних конференціях «GRANT-2015» (2015, Київ) «Відкриті семантичні технології для інлелектуальних 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-ій 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.

РОЗДІЛ 1
ДЖЕРЕЛА НАУКОВОЇ ТА НАУКОМЕТРИЧНОЇ ІНФОРМАЦІЇ. СПОСОБИ ІНТЕРПРЕТАЦІЇ ДАНИХ

1.1 Доступ до науково-технічної інформації – регламент, провідні ресурси

Доступ до науково-технічної інформації є одним із визначних факторів актуальності даної інформації з огляду на загальносвітову тенденцію щодо доступності інформації.

Також формується загальнодержавна, в деяких напрямках життєдіяльності людини, політика щодо відкритості інформації, приватності та захисту персональних даних та право на неоприлюднення тієї чи іншої інформації.

Одним із яскравих прикладів розвитку такої політики є запровадження нових загальних норм із захисту даних в Європейському Союзі [1]. Запровадження нових норм так само передбачає змін у політиках поширення інформації в соціальних мережах, наприклад відповідні анонси було оприлюднено для користувачів Twitter [2]. Також відповідні зміни відбулися для користувачів одного з найбільших сервісів мережі інтернет – Google [3].

В той же час необхідно звернути увагу на відповідні політики щодо відкритого доступу. Кожна країна визначає свої обмеження щодо відкритих даних і способу їх оприлюднення. Для України набори даних, що мають бути оприлюднені у відкритому доступі визначаються Постановою Кабінету Міністрів України від 21 жовтня 2015 року №835 «Про затвердження Положення про набори даних, які підлягають оприлюдненню у формі відкритих даних» [4]. Цим документом визначаються всі аспекти оприлюднення відкритих даних, зокрема терміни, учасники, паспорти наборів, тощо.

Значне місце у просторі відкритих даних та порядку їх регулювання займає науково-технічна інформація інформація: наукові публікації та наукові данні.

За загальним формулюванням відкритий доступ передбачає вільну безкоштовну і безперешкодну он-лайн публікацію результатів досліджень із

можливістю застосування ліцензій Creative Commons (https://creativecommons.org/) до розміщеного матеріалу [5].

Зокрема грантові програми різноманітних напрямків і способів підтримки наукових досліджень передбачають публікацію результатів досліджень та розміщення у відкритому доступі даних, отриманих за результатами проведення дослідження. Такі вимоги чітко окреслені у вимогах участі в програмі Горизонт-2020 – найбільшій програмі Європейського Союзу з досліджень та інновацій, і передбачає дотримання цих вимог [6]. Також відповідно до робочої програми нової рамкової програми з досліджень на інновацій Європейського Союзу «Горизонт Європа» на 2021-2027 роки відкрита наука є осноним принципом роботи програми і вимагатиме відкритий доступ до публікацій, даних, та планів керування даними [7].

Також можна говорити про важливість інформації про досягнення вітчизняних вчених черех призму конкурсної діяльності [8].

В той же час відкритість тієї чи іншої наукової інформації визначається регуляторними нормами грантових програм, видавництв, університетів, організацій, що опікуються цією інформацією та Законом України про національну програму інформатизації [9].

На сьогодні в мережі Інтернет налічується значна кількість ресурсів, що передбачають доступ як відкритий так і передплачений до наукової інформації та наукових даних. Відповідно кожен ресурс передбачає власні політики щодо публікацій та умов розміщення інформаційних масивів або наборів даних на ресурсах, що є фактором, який має бути врахований при побудові інформаційних систем та виокремленні нових масивів інформації.

При розгляді наукометричних ресурсів варто зазначити найбільш відомі і визнані в світі – Web of Science, Scopus та Google Scholar. Важливо підкреслити, що лише Google Scholar є ресурсом, що знаходиться у відкритому доступі і не передбачає передплати або оплати за доступ. Дані, які представляє інформаційний ресурс оновлюються швидко і в такий спосіб надає користувачеві актуальну інформацію. Ці ресурси є найбільш авторитетними в світі, дані цих ресурсів

враховуються при аналізі та оцінці наукової діяльності як окремих науковців, так і наукових колективів і організацій.

В той же час компанія-провайдер науково-технічної інформації Elsevier надає можливість отримати окрему інформацію та розраховані наукометричні показники, надаючи відкритий доступ до даних за рахунок джерел – Science Direct та Scimago Journal&Country Rank.

Важливим є зазначення принципу побудови ресурсів наукової інформації – яка саме набори даних є доступними користувачеві для подальшого розуміння необхідності побудови нових моделей для реалізації актуального інструментарію на базі існуючих ресурсів.

За рахунок реалізації задач формування нових масивів інформації на базі аналізу ресурсів наукометричної інформації постає питання отримання, структуризації та візуалізації даних шляхом сканування ресурсів повнотекстової інформації відритого доступу. Також одним із елементів співставлення традиційних показників та аналітичних висновків можуть бути дані енциклопедичних ресурсів, що є складовими наукової інформації. Зазначені ресурси також містять необхідне підгрунтя для проведення наукометричного аналізу та формування нових показників.

Підходи щодо залучення наукометрії в інформаційних системах для оцінки наукової діляьності є актуальним питанням та зображено зокрема в роботах Хассана та Лоебеке [10].

При розгляді ресурсів наукової інформації крім наукометричних ресурсів також необхідно взяти до уваги архіви електронних публікацій та репозитарії що представлені в авторитетному світовому рейтингу Webometrics [11], так і вузькоспеціалізовані ресурси. Серед розглянутих систем однією із найбільш популярних, авторитетних та такою, що містить велику кількість документів – наукових публікації є arXiv - ресурс препринтів Корнуельської бібліотеки.

В першу чергу варто розглянути наступні ресурси:

- наукометрична та пошукова система наукової літератури – Google Scholar.

- найбільша он-лайн енциклопедія Вікіпедія.
- архів електронних препринтів arXiv.

З огляду на поставлені в роботі задачі було опрацьовано літературу із питань – розробки надбудов для зазначених систем, що розширюють можливості інформаційних ресурсів і дають змогу користувачеві отримати нові масиви інформації на їх базі.

1.2 Попередні дослідження щодо аналізу даних ресурсів наукової та наукометричної інформації

На сьогодні науковців в першу чергу цікавить питання порівняння показників провідних наукометричних ресурсів, зокрема – співставлення показників системи відкритого доступу Google Scholar та передплачуваних ресурсів – Scopus та Web of Science.

За останні 10 років питанню порівняння систем було присвячено наступні роботи:

- порівняння систем PubMed, Scopus, Web of Science та Google Scholar [12] – роботу присвячено порівнянню корисності інформації, що надають системи за заданим концептом із біологічного спрямування з огляду на актуальність зазначених ресурсів, які надають вичерпну інформацію щодо цитованості тих чи інших робіт, що було представлено за результатами пошуку;
- переваги та недоліки системи Google Scholar було опрацьовано та представлено вченим університету Гаваїв, Пітером Яско [13], який прослідкував історію та пошукові результати із дати створення ресурсу – 2004 рік, та у висновках зазначив, що за усунення недоліків у роботі системи Google Scholar незабаром стане зручним безкоштовним інструментом для пошуку наукової інформації;
- порівняння індексу Гірша, що обраховується системами Web of Science, Scopus та Google Scholar [14]. У своїй роботі автор досліджує не тільки

обсяг наукових джерел, якими оперуються системи для обрахунку індексу Гірша, а й способи та підходи до реалізації обрахунку індексу Гірша;

- в роботі дослідників Університету Мельбурна [15] Google Scholar розглядається як зручна альтернатива 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, не зважаючи на їх бюджети і спроможність передплачувати доступ до наукометричних ресурсів;
- порівняння цитувань в журналах з медицини – мета дослідження науковців Університету Торонто [15]. В публікації зазначено основні методи та порівняльний аналіз цитувань, що представлені ресурсами Web of Science, Scopus та Google Scholar;

З огляду на те, що Вікіпедія на сьогодні є найбільш популярною енциклопедією в мережі інтернет, навколо існує багато думок щодо коретності та перевіреності інформації, що представлена в системі. Дослідження, що стосуються Вікіпедія в більшості пов’язані із посиланнями на ресурс і відповідно, на які ресурси є посилання в статтях Вікіпедії, що зайвий раз може доводити актуальність та достовірність представленої інформації з огляду на можливість редагування публікацій всіма користувачами мережі інтернет. Також багато робіт присвячено обробці гіперпосилань у статтях Вікіпедія для побудови семантичних зв’язків та відповідності та доповненості понять, висвітлених у публікаціях Вікіпедія.

Серед робіт, що за останні 10 років було присвячено зазначеній проблематиці наступні:

- питання цитування Вікіпедія – є актуальним і робота колективу дослідників Університетів Оттави та Торонто присвячено виокремленню та аналізу публікацій з напрямку медицини, що посилаються на Вікіпедію. Результатом якого стало зазначення, що на статті у Вікіпедії посилаються

не лише наукові журнали з малим імпакт-фактором, а й журнали із великим індексом цитування, з яких порядку 30% посилань – є визначення певних понять [17];

- шляхом аналізу публікацій за допомогою реферативних баз даних Web of Science та Scopus було ідентифіковано географічне розташування, відповідно авторів, інституції, наукові напрямки із найбільшим цитуванням Вікіпедії [18];
- одна із робіт [19], що присвячена вивченню джерел публікацій Вікіпедії досліджує кількість наукових публікацій, що водночас статтями з Вікіпедії чи є їх складовою. А саме – що це за публікації, з яких галузей знань та наукових напрямків із представленою у публікації відповідною статистикою.
- важливим питанням, що корелюється із запропонованими у роботі питаннями побудови онтології для заданого поняття на базі енциклопедичного ресурсу Вікіпедія висвітлено у кількох роботах. Один із підходів, що пропонується [20] – використовувати імовірнісну модель, що враховує кількість разів, що зустрічається об’єкт у статтях, і подальше представлення у вигляді мережі Баєса;
- формування рефератів та анотацій для статей є актуальним питанням і достатньо багато програмних продуктів присвячено вирішенню цієї задачі [21]. Проте нові підходи до реалізації подібних задач висвітлено в науковій літературі [22], і базуються на використанні гіперпосилань.
- визначення тематики документів за допомогою визначення ваги слів, що є гіперпосиланнями в публікаціях Вікіпедія [23];
- на базі гіперпосилань можлива побудова наборів текстів, що будується за допомогою сегментації слів у реченні [24]. Підхід реалізовано на базі сучасного івриту, але може бути застосований для інших східних мов;
- кластеризація документів Вікіпедіїї в такий спосіб, щоб у подальшому можна було відображати документи в концепції та підрозділах Вікіпедії [25];

Одним із підходів, що є подібним до запропонованого у дисертаційному дослідженні є інклюзивний підхід, в порівнянні із селективним[26], для створення медичної онтології за рахунок інклюзії всіх назв статей Вікіпедії, як концептів та всіх існуючих посилань як потенційних зв'язків. Метою підходу є створення направленого немаркованого графу, що імітує структуру посилань.

З огляду на те, що Вікіпедія є відкритим ресурсом, що дозволяє не тільки обробляти текстову інформацію, яку містить енциклопедія, а й формувати надбудови та аналітичні висновки з огляду на доступність використання програмних додатків – інтерфейсу прикладних програм, питання побудови онтології є популярним серед науковців і присвячені також:

- використанню інформації з Вікіпедії для побудови гео-онтології [27] – виокрмлення данних від укрупнених – континентів до найдрібніших – поселення. Вікіпедія у дослідженні виступає джерелом даних для розробки інструментарію для обробки тексту природньою мовою. Також дослідження реалізовано за рахунок сканування сторінок Вікіпедії, що містять інформацію про географічне розташування об'єктів і подальше порівняння із реальним даними, а також побудова та відображення зв'язків між об'єктами в рамках онтології;
- побудові великих мереж для онтології персоналій [28], що передбачає екстракцію понять, що представляють собою власні імена людей за рахунок використання класифікатору машинного навчання, виокремлення «is-a» взаємозв'язків, збір назв статей Вікіпедії, що представляють собою власні імена людей;
- застосуванню Вікіпедії, як онтології для опису документів [29]. Дослідження спрямоване на виявленню концептів, що відносяться до наборів документів, за рахунок використання текстів статей Вікіпедії та гіперпосилань.

Окрім аналізу гіперпосилань в текстах статей Вікіпедії дуже важливим аспектом є посилання зовнішні. Традиційно для енциклопедичного ресурсу

Вікіпедія, кожна стаття містить гіперпосилання на інші статті Вікіпедії, але в той же час є посилання і зовнішні, що направляють користувача на ресурси поза Вікіпедією. Робота дослідників Університету Патраса та Іонійського Університету (Греція) [30] присвячена безпосередньо якісній оцінці зовнішніх посилань в Вікіпедії. Дослідження передбачає вивчення зовнішніх посилань Вікіпедії за рахунок оцінки ступеню їх відповідності призначенню – формування повного переліку джерел інформації про зміст статті. Важливим є вивчення розподілу зовнішніх посилань в статтях Вікіпедії та доведення їх цінності, що в свою чергу підкреслює цінність інформації, що містять статті ресурсу.

Також дуже важливо звернути увагу на суперечливі роботи, що заперечують використання інформації з енциклопедичного ресурсу Вікіпедія для опису дослідження у наукових публікаціях, зокрема – цитування у наукових працях [31]. Адже Вікіпедія не може розглядатися як незалежне джерело інформації і радше звертатися до першоджерел, які містяться у посиланнях статей Вікіпедії. В дослідженні зображено, зо серед 1400 наукових публікацій, що містили посилання на статті Вікіпедії – лише 4% були доречними. І відповідно постає питання щодо актуальності та доцільності використання інформації Вікіпедії для побудови додаткових масивів інформації.

В той же час окрім наведених вище робіт, що зображують аналітику щодо посилань з різних галузей знань на статті Вікіпедії, є такі, що розглядають причини звернення вчених до енциклопедичного ресурсу для формування власних наукових текстів [32]. В статті зазначається обсяг предметних областей, публікаційна активність вчених – представників визначених предметних областей, звертається до Вікіпедії як до джерела інформації, і підкреслюється, яка саме інформація цитується – факти, визначення, статистика, тощо.

Однак Вікіпедія розглядається і як корпус для вилучення знань [33], адже з огляду на те, що ресурс є енциклопедичний – на ньому представлено багатий обсяг концептів з усіх галузей знань і дозволяє проводити аналіз інформації та відповідно проводити структуризацію і пошук точок дотику між ними.

Ресурси відкритого доступу однозначно мають перевагу перед передплаченими ресурсами і не тільки з огляду на доступність інформації, а й можливості виокремлення даних та отримання нової інформації, розробки додатків (відповідно до ліцензій відповідного ресурсу) і надавати користувачеві більш широкий спектр інформаційних наборів для подальшої її обробки чи аналізу.

Одним із ресурсів, на базі якого в рамках дисертаційного дослідження було побудовано нові масиви інформації є архів препринтів arXiv.

Ресурс переважно цікавить науковців з точки зору розміщення своїх публікаційних доробків, адже це – можливість до публікації у науковому журналі – он-лайн чи друкованому, відповідно до політик видання зробити свої наукові відкриття доступними.

В той же час цікавість науковців щодо ресурсу препринтів є з наступних напрямків:

- з огляду на відкритість ресурсу і актуальність робіт, що є розміщеними в архіві поряд із опцією щодо отримання повідомлення про нові публікації з тематики, що є у сфері зацікавленості користувача дослідити важливим питанням дослідників у роботі [34] стало питання кількості завантажень та цитувань робіт з arXiv у галузі астрофізики;
- порівняльний аналіз публікацій з напрямку астрофізики на прикладі ресурсів arXiv, Scopus та Mendeley [35];
- вплив можливостей, що надає ресурс на збільшення кількості посилань на публікації з математики безпосередньо на arXiv та відповідному зменшенню завантажень статей із сайтів видавців [36];
- розширення можливостей ресурсу за рахунок додавання опцій проведення та написання на сайті відгуків експертів – проведення рецензування публікацій архіву за допомогою інструментарію системи [37].

Цілком закономірним є увага дослідників до напрямку астрофізики адже із самого початку архів препринтів arXiv було розраховано на розміщення статей із фізики, відповідно кількість статей – матеріалу для дослідження найбільше.

1.3 Інформаційні ресурси наукової інформації

При розгляді джерел наукової інформації необхідно в першу чергу сформулювати визначення інформації. Під інформацією будемо розуміти певну сукупність даних, що набуваються із навколишнього середовища шляхом спостереження та інших способів збору інформації, і так само ці дані можуть передаватися у навколишнє середовище в різний спосіб, або сукупність даних може зберігатися всередині визначеної системи. Інформація може бути представлена у різний спосіб. Одними з найбільш важливих властивостей інформації можна вважати наступні: об'єктивність та суб'єктивність, доступність, актуальність, цінність, повнота, достовірність, адекватність. Саме із урахуванням цих властивостей інформації можуть бути спроектовані щодо наукової інформації та їх базі можуть бути сформовані критерії для оцінки ресурсів наукової інформації. В свою чергу визначення наукової інформації може бути сформульовано у наступний спосіб – це інформація, шо є логічною та її можна отримати в процесі пізнання, це та інформація, що здатна адекватно відображати закономірності об'єктивного світу, використовується для розвитку наукових знань та вдосконалення процесу пізнання. Однією із важливих ознак наукової інформації можна вважати сукупність документів, що містять опис та обґрунтування досягнень, що були отримані в результаті науково-дослідної діяльності. В такий спосіб можна зазначити такі основні джерела наукової інформації:

- монографія – праця в конкретній вузькій галузі знань, що ґрунтовно викладає вичерпний опис, результати та перспективи розвитку окресленого в праці питання. Як правило містить не менше 5-6 друкованих аркушів. Така наукова праця передбачає рецензування щонайменше двома рецензентами. Може бути складена одним або кількома авторами;
- наукові періодичні видання – це журнали та інші періодичні видання, що представляють собою сукупність публікацій із конкретного наукового напрямку або проблеми, сукупність публікацій з різних галузей знань. Наукові публікації, що надруковані у наукових періодичних виданнях містять викладення процесу наукового дослідження і основних

результатів. Як правило всі матеріали, що надруковані у наукових періодичних виданнях проходять процедуру рецензування щонайменше двома рецензентами.

- стандарти – сукупність нормативних документів, що є вимогами до реалізації наукових досліджень та представленні результатів та наукової продукції.

Тож інформаційні ресурси наукової інформації – це сукупність джерел наукової інформації, інформації про зазначені джерела та основні характеристики та параметри, що їх описують.

Інформаційні ресурси в їх традиційному сприйнятті можуть бути як у вигляді друкованих видань, так і актуальних на сьогодні он-лайн інформаційно-аналітичних систем. Інформаційна система (ІС) – ситема, що реалізує інформаційну модель визначеної предметної галузі діяльності людини. Інформаційна система має забезпечувати отримання, пошук, передачу та перетворення інформації. Як правило представляє собою сукупність взаємопов’язаних апаратно-програмних засобів, що призначені для автоматизації обробки даних з різних джерел, що в поадльшому зберігаються, перетворюються та надаються користувачеві. Інформаційна система складається з джерела інформації, апаратної частини, програмного забезпечення, споживачів інформації, персоналу, що обслуговує систему [38].

Основними задачами ІС є належне збереження та обробка інформації. В контексті наукової інформації – ІС наукової інформації спрямовані в першу чергу на надання наукової інформації, джерел наукової інформації, їх характеристик. Інформація, що надається зазначеними ресурсами орієнтована на широке коло користувачів з огляду на потребу у такій інформації не тільки фахівців вузького наукового напрямку, а й фахівців інших спеціальностей для задоволення пізнавальних потреб. З огляду на реалізацію тих чи інших потреб необхідно окреслити перелік даних та масивів інформації, що можуть містити інформаційно-аналітичні системи наукової інформації:

- наукова публікація і відповідні її характеристики (предметна область, науковий напрямок, назва публікації, автор, ключові слова, реферат (абстракт, анотація), повний текст публікації), а також посилання на повний текст публікації у випадку скороченого представлення інформації про публікацію, в залежності від типу публікації, можливе зазначення також – назви наукового періодичного видання, назва заходу, на якому було представлено результати дослідження, назва видавництва;
- наукометричні показники (імпакт-фактор видання, в якому опубліковано наукову працю, індекс Гірша автора публікації, кількість цитувань наукової праці);
- наукові дані – безпосередньо результати наукових досліджень, представлені згідно вимог до представлення результатів вимірювань, обчислення, тощо. Поряд із науковими даними завжди містить інформація про дослідження в рамках якого було отримано представлені результати, зазвичай посилання на грантову підтримку реалізованого проекту;
- текстова інформація, що представляє собою набір текстових документів: визначень, текстової інформації про ті чи інші наукові концепти.

Згідно із класифікацією інформаційних систем можна зазначити, що інформаційні системи наукової інформації:

- за характером організації пошуку інформації – інформаційно-пошукові та інформаційно-довідкові;
- за характером функціональності – багатофункціональні;
- за масштабністю реалізації – локальні, регіональні і глобальні;
- за ступенем автоматизації – автоматичні;
- за характером обробки даних – інформаційно-довідкові або інформаційно-пошукові ІС, в яких немає складних алгоритмів обробки даних, а метою системи є пошук і видача інформації в зручному вигляді;
- за сферою застосування – інформаційно-довідкова система, що охоплює різні предметні області.

Згідно до вимог, що висуваються до побудови інформаційних систем варто

Відмітити основні вимоги, яких мають дотримуватися інформаційні системи для організації корисного діалогу із користувачем:

- достовірність – система не має містити помилкових даних;
- релевантність – забезпечення впевненості в отриманні очікуваної інформації;
- повнота – наявність повного набору даних, що цікавлять дослідника;
- порівнянність – можливість порівняння інформації із іншими джерелами (у випадку наукометричної інформації – можливість порівняти дані ресурсів, що обчислюють та представляють одні і ті самі показники);
- цілісність – коректність представлення та збереження інформації для повного представлення її користувачеві – без змін.

Для формування переліку ресурсів наукової інформації та реалізації систематизації зазначених систем надамо визначення поняття систематизація. Під систематизацією розуміємо виокремлення спільних ознак систем глобальної мережі Інтернет, що містять наукову інформацію в різних наборах, способах та обсягах її представлення, для формування єдиної системи ресурсів наукової інформації.

Необхідно також зауважити, що розглядаючи ресурси наукової інформації необхідно враховувати, що інформація про грантові пропозиції та результати реалізації проектів за грантової підтримки також є частиною наукової інформації, і способи організації ресурсів фондуючих організацій [39] також можуть розглядатися в цьому контексті. І критерії та підходи для визначення показників успішності вчених будуються саме на базі таких ресурсів [40, 41, 42].

1.4 Систематизація ресурсів наукової інформації

Згідно аналізу останніх досліджень та публікацій в англомовному сегменті наукових публікацій за останні 3 роки увагу авторів та дослідників було зосереджено на таких питаннях щодо систематизації ресурсів наукової інформації та способів аналізу ресурсів наукової інформації:

- підходи до формування інтелектуальних наукових інтернет-ресурсів [39];
- дослідження інформаційних ресурсів, як один із нових наукових напрямків, що передбачає оцінку наукових ресурсів [43];
- нові підходи щодо систематизації досліджень, що базується на підході «Big Data» [44]
- формування кількісного та якісного дискурсу щодо використання даних, як дослідницької техніки [45];
- формування систематизованої структури оцінки ефективності користування наукових ресурсів [46];
- окреслення підходів щодо формування суспільства даних з огляду на спрямованість сьогодення щодо використання та обробки великих масивів даних [47].

В україномовному сегменті тематика наукових робіт зосереджена в публікаціях напрямку інформаційно-бібліотечної діяльності, зокрема в них підняті такі питання:

- формування та використання електронних ресурсів наукової та освітньої інформації [48];
- питання формування, впорядкування та управління в контексті бібліотечних інформаційних ресурсів, що містять посилання на ресурси наукової інформації, тощо [49];
- особливості розробки критеріїв оцінювання електронних ресурсів освітнього напрямку [50];
- проблема захисту наукових ресурсів [51].

Таким чином можна зауважити, що робіт щодо систематизації або формування підходів до систематизації наукових ресурсів та визначення критеріїв для аналізу та оцінки ресурсів наукової інформації з точки зору доступності і контенту не провадилось. В той же час матеріали досліджень, що присвячені способам розгляду наукових та освітніх ресурсів стали підґрунтям у формуванні

підходів до способів систематизації зазначених систем та способам визначення критеріїв для подальшої роботи.

1.5 Визначення характерних ознак інформаційних систем наукової інформації

Для реалізації поставленої задачі згідно наданого визначення, постає необхідність визначення характерних ознак систем наукової інформації, що відповідають критеріям для оцінки цих систем.

Питання аналізу ресурсів, або визначення критеріїв для аналізу ресурсів зосереджено на особливостях розробки критеріїв для оцінювання електронних освітніх ресурсів з метою підвищення якості знань, що набуває користувач.

З огляду на актуальність поставленої задачі та наявності в глобальних мережах великої кількості ресурсів наукової інформації, важливим постає питання систематизації таких ресурсів та визначення основних характеристик з огляду на наповнення та способу розробки систем – достатність інформації та врахування орієнтованості щодо користувача. Визначення критеріїв для аналізу та оцінки зазначених ресурсів надає можливість також здійснити опис ресурсів наукової інформації з точки зору її доступності та наповнення дозволить не тільки орієнтуватися у масиві ресурсів, що наразі є доступними користувачам мережі Інтернет, а й вдосконалювати існуючі системи, з огляду на вимоги щодо розробки інформаційних систем та користувацького інтерфейсу зокрема, та визначати перелік даних відповідно до задачі тієї чи іншої системи.

Основними завданнями дослідження було розробити модель, що дозволить виокремити низку показників, за якими буде проводитись аналіз ресурсів наукової інформації для проведення оцінки ресурсу з огляду на:

- відкритість даних;
- спосіб представлення інформації;
- варіанти наборів даних, що представлені на ресурсі;

Для реалізації поставленої задачі було проведено та реалізовано наступні

етапи:

- запропоновано представлення наукової інформації у вигляді мережі, елементами якої є користувачі інформації, способи представлення наукової інформації, доступу до неї, деталізація контенту;
- експертним шляхом було виокремлено ресурси, що містять наукову інформацію (результати досліджень – наукові дані, препринти наукових публікацій, наукові публікації, реферативна інформація щодо наукових публікацій, наукометричні показники). Проаналізовано спосіб представлення інформації та способи пошуку інформації в рамках ресурсу;
- визначено спільні ознаки, як у контенті ресурсів різного спрямування так і способі представлення інформації;
- відповідно до аналізу зв’язків мережі виокремлено критерії для опису та аналізу ресурсів наукової інформації.

Запропоновано представлення компонентів, що характеризують наукову інформацію в контексті її представлення в мережі Інтернет, споживачів цієї інформації, складових наукової інформації та зв’язків між цими компонентами у вигляді графа (рис. 1.1).

Рисунок 1.1 – Основні компоненти наукової інформації, в контексті її представлення в мережі Інтернет

Схема містить наступні основні елементи:

- набір систем (енциклопедичні ресурси, ресурси видавництв, репозитарії публікацій, наукометричні ресурси, репозитарії наукових даних);
- споживачі інформації (переічний користувач (користувач-ненауковець), студент, аспірант, науковець);
- мета доступу до інформації (пошук інформації за темою, підготовка до зайнять, написання наукової роботи, написання публікації, підготовка грантової пропозиції);
- пошук інформації (традиційні пошукові системи, спеціалізовані пошукові системи).

За рахунок реалізації поставленої задачі, зокрема визначення критеріїв для опису ресурсів, мережа може бути доповнена зв'язками між компонентами даних, що містять системи наукової інформації та безпосередньо самими системами, як джерелами наукової інформації.

1.6 Представлення наукової інформації на рексурсах мережі Інтернет

Запропонована схема є описовою моделлю для зображення наукової інформації з точки зору виокремлення її основних компонентів та в контексті представлення її у мережі Інтернет.

З огляду на те, що після проведення аналізу систем наукової інформації запропоновані критерії можуть бути присутні або ні, тобто кожен з критеріїв може мати значення «0» чи «1», то дана модель може бути формалізована у вигляді матриці, що має значення «0» та «1» (Таблиця 1.1).

Кожен критерій може бути представлений у вигляді множини змінних:

$$K = \{k1, k2, k3, \ldots, kn\},$$

а множина інформаційних систем наукової інформації, як множина у вигляді:

$$S = \{s1, s2, s3, \ldots, sm\}.$$

Якщо припустити, що всі критерії мають однакову вагу, то за рахунок оцінки таблиці, що може бути складена за результатами розгляду систем – можна

сформувати рейтинг систем наукової інформації та виокремити сильні та слабкі сторони.

Таблиця 1.1 – Логічне представлення відповідності критеріїв для аналізу наукової інформації – конкретним інформаційним системам.

	k1	k2	k3	…	kn
s1	0	1	0	…	0
s2	0	1	0	…	1
s3	0	1	1	…	0
…	…	…	…	…	…
sm	1	0	1	…	1

В той же час важливим є виокремлення типів систем наукової інформації і відповідність їх запропонованим критеріям.

1.7 Типи ресурсів наукової інформації

Для реалізації поставленої задачі запропоновано наступні типи ресурсів наукової інформації, що можуть бути розглянуті і відповідно тип системи – виступає пошуковим словом для подальшого експертного пошуку та визначення систем для подальшого дослідження:

- наукометричні ресурси – на сьогодні є актуальним джерелом отримання інформації не тільки про публікації та видання за всіма науковими напрямками, а й ресурсом, що надає інформацію про наукометричні показники – такі як імпакт-фактор наукового видання, індекс Гірша, цитування, самоцитування, тощо.

  Серед провідних світових наукометричних ресурсів розглядатимемо наступні: Web of Science, Scopus, Google Scholar.
- репозитарії наукових текстів та публікацій – є інформаційними системами, що представляють собою систему, що зберігає великий обсяг наукових текстів: опублікованих праць, препринтів, презентацій та інших

наукових документів. Репозитарії є системами локальними, якщо ми розглядаємо такі системи з точки зору контенту, що належить, як правило, науковцям, аспірантам, здобувачам та студентам одного навчального закладу чи наукової установи. В той же час розміщення таких систем в мережі Інтернет передбачає вільний доступ до розміщених документів для всіх користувачів глобальної мережі. Також репозитарії наукових текстів можуть бути представлені у вигляді систем, що зберігають та представляють інформацію щодо наукових текстів конкретної предметної області.

Для аналізу репозитаріїв наукових текстів було обрано наступні ресурси:

Research and Publications Archive, Research Publications Repository, Joint Research Centre Publications Archive.

Архіви препринтів та публікацій – інформаційні системи, що дозволяють оприлюднити результати своїх досліджень до подання для друку до наукових періодичних видань. Кожен автор розміщує тексти своїх досліджень, що дозволяє до виходу публікації (що вимагає тривалого часу) закріпити за собою авторство щодо результатів дослідження та представити їх на розсуд наукового загалу і надати можливість дослідникам дізнатися про останні новинки в галузі. Як правило кожен препринт проходить процедуру рецензування для уникнення публікації вже існуючих матеріалів. Для проведення аналізу інформаційних систем – архівів препринтів та публікацій було обрано два найбільші ресурси мережі Інтернет – архів препринтів arXiv, що охоплює кілька наукових напрямків та архів публікацій Zenodo. В той же час варто відмітити, що на сьогодні достатньо широко розвивається мережа ресурсів препринтів з окремих вузьких галузей знань та окремих галузей знань, такі як – bioRxiv (архів препринтів з біології), PeerJ Prerpint Archive (архів препринтів з хімії, математики та комп'ютерних наук), ASAPBio (архів препринтів з біології). Також розповсюджене створення архівів препринтів в рамках репозитаріїв окремих університетів.

Ресурси наукових видань/видавництв наукової літератури – безпосередньо веб-сторінки видавництв наукової літератури та періодичних наукових видань. Ресурси містять інформацію про видання, порядок прийняття о розгляду матеріалів, і відповідно до політик видавництва/видання – повні тексти публікацій (архів номерів) або реферативна інформація про них. Серед видавництв та видань було обрано наступні: Springer, Elsevier, International Organization of Scientific Research, Scientific Research, International Journal of Computer, Open Science Journal.

Репозитарії даних наукових результатів – інформаційні ресурси, що передбачають доступ до наукових результатів – масивів даних, що були отримані безпосередньо в результаті проведення досліджень. Як правило – це є результати роботи над проектами в рамках грантової підтримки та результати роботи в рамках колаборацій. Для реалізації аналізу було обрано Інтегрований ресурс доступу до наукової інформації та публікацій OpenAire та інформаційний сервіс ЦЕРН – CERN Scientific Information Service.

Пошукові ресурси – провайдери доступу до ресурсів наукових та освітніх даних є цільовою пошуковою системою, що спрямована на пошук наукової інформації. Варто зауважити, що поряд із пошуковими ресурсами деякі системи, що було віднесено до наукометричних – Google Scholar – також є пошуковою системою. Для дослідження було розглянуто ресурс Registry of Research Data Repositories.

Окрім розглянутих ресурсів варто звернути увагу на ресурси – ідентифікатори авторів, що передбачають зазначення інформації про автора, його публікаційні доробки, афіліацію, а також ключові слова. Серед найбільш популярних ресурсів – ORCID та ResearchersID.

Також – соціальні мережі науковців – один із популярних та ефективних інструментів, що використовуються науковцями для обговорення своїх результатів, формування дослідницьких груп. Мова йде безпосередньо про соціальні мережі для науковців – ресурси, що дозволяють науковцям створювати власні профілі, розміщувати як опубліковані (відповідно до політик видань) так і підготовлених до друку публікацій, зазначати інформацію про себе та отримати

можливість винести на розгляд наукового загалу результати своїх досліджень. Одним із провідних світових ресурсів – соціальна мережа для науковців ResearchGate.

ResearchGate – соціальна мережа науковців, що дозволяє розміщувати власні наукові доробки – публікації, наукові дані, для подальшого поширення іншими ресурсами та обговорення із колегами.

Ресурс є відкритим для користувачів Інтернет, але доступ до всіх даних передбачає реєстрацію, або синхронізацією із профілем в соціальній мережі LinkedIn або Facebook. Безпосередня реєстрація можлива лише при зазначенні інституціональної електронної адреси.

З огляду на ключові компоненти, що характеризують систему було виокремлено перелік компонентів, що є притаманними для систем, які містять наукову інформацію, і відповідно було проаналізовано ресурси.

Схему компонентів було визначено з огляду на елементи, що характеризують наукову публікацію із додаванням елементу повного тексту або посилання на повний текст, що є невід'ємним атрибутом ресурсів, що містять інформацію про наукову публікацію. Зазначені елементи є групою ознак, що дозволяють визначити варіанти наборів даних, що містить той чи інший ресурс.

Також при формуванні переліку компонент було враховано складові інтерфейсу користувача (рисунок 1.2), що містить інформацію про способи вводу і виведення інформації, засоби її відображення та візуалізацію [52], і також який є невід'ємним елементом схеми організації діалогу з використанням моделі користувача.

1.8 Модель користувача

Модель користувача передбачає розгляд станів, які може набувати користувач при користуванні системою. Серед таких станів – порядок дій, що виконує користувач при роботі з системою. Це так звана прогнозована модель користувача, що може бути побудована на базі марківських ланцюгів. Дана модель базується на припущені про поведінку користувача: ймовірність кожних наступних

кроків залежно від попередніх. Саме тому доречно використання марківських ланцюгів та прихованих марківських моделей. Перевага використання такого підходу до формування моделі користувача в тому, що розмір марківських ланцюгів збільшується експоненціально залежно від росту порядку, а в зазначеній моделі такий порядок не є великим.

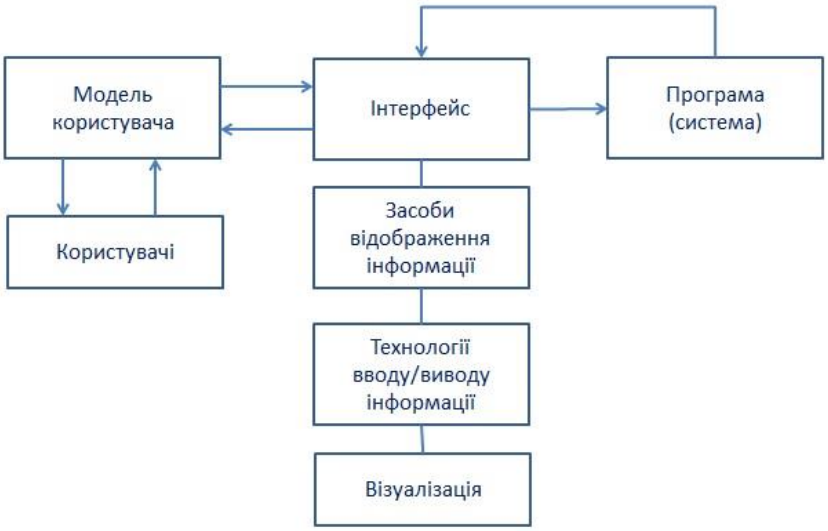


Рисунок 1.2 – Схема організації діалогу користувача з системою

Також варто зазначити, що поведінка кожного наступного користувача в тій чи іншій мірі повторює поведінку попереднього, адже мета звернення до ресурсу є спільною – пошук наукової інформації за визначеними ключовими словами/словосполученнями.

На рисунку 1.3 представлено модель поведінки користувача при роботі із інформаційною системою наукової інформації (наукометрична система, архів препринтів, репозитарій).

Серед послідовності дій, що виконує користувач при роботі із системою науковою інформації виокремлено наступні:

- відкриття веб-сторінки системи;
- пошук інформації за ключовим словом або словосполученням;
- перегляд отриманої інформації;
- уточнення пошуку;
- повернення на головну сторінку;

- перегляд інформації про публікацію/видання;
- завантаження тексту публікації.

Рисунок 1.3 – Модель користувача при роботі з інформаційною системою наукової інформації.

Висновки до розділу 1.

З огляду на те, що увага наукового загалу є виключно до кількісних показників та питанню цитованості в окремих галузях знань, робота запропонована у дисертаційному дослідженні є актуальною з точки зору пошуку міждисциплінарної складової за рахунок аналізу наявних предметних областей та розміщених за ними публікацій згідно заданого концепту.

Широкий спектр ресурсів, що розглянутий в роботі дозволяє:

- порівняти інформаційний контент та способи представлення та доступу до інформації в ресурсах відкритого доступу та передплачених ресурсах;
- виокремити основні характеристики наукової публікації і відповідні дані та наукометричні показники, що обраховуються на базі цієї інформації;
- оцінити підходи до побудови систем наукових даних;

- оцінити можливості програмної реалізації розроблених моделей для отримання та подальшої роботи із даними ресурсів наукової інформації;
- вивчити існуючі та запропонувати нові підходи до візуалізації отриманих результатів;
- визначити ліцензії та інші обмеження щодо використання та обробки даних, представлених на ресурсі.

За рахунок огляду літератури щодо опису джерел наукової та наукометричної інформації та підходів до інтерпретації відповідних даних дає можливість підкреслити важливість інформації, адже переважно цікавіть наукового загалу, що цікавиться питаннями наукометрії та інформаційних технологій та даних на базі цих ресурсів зосереджена на порівнянні даних, що надають системи, способи розширення можливостей систем для проведення подальших досліджень, а не надання користувачеві додаткових масивів інформації та оптимізації роботи із даними.

РОЗДІЛ 2

ПІДХОДИ ДО ФОРМУВАННЯ КРИТЕРІЇВ ОЦІНКИ РЕСУРСІВ НАУКОВОЇ ТА НАУКОМЕТРИЧНОЇ ІНФОРМАЦІЇ

2.1 Аналіз ресурсів наукової та наукометричної інформації

Отже за рахунок аналізу послідовності дій користувача та фіксації відгуків системи можна виокремити ряд критеріїв, які запропоновано для аналізу систем наукової інформації. Запити користувача формуються відповідно до характерних ознак публікації, зокрема ключових слів, за якими в подальшому користувач формує картину повноти та завершеності пошуку інформації в системі [53].

Також при формуванні критеріїв дуже важливо зважати на те, в який спосіб буде представлена інформація, яку запитує користувач, і також фактор можливості збереження її на комп'ютері користувача для подальшого її використання та всі опції щодо наборів даних, що є доступним для користувача [54].

Розглянемо можливі способи відображення на збереження інформації, що надає інформаційна система наукової інформації:

- візуальна інформація, завантаження та збереження якої можливе тільки за рахунок використання screen-shot;
- візуальна інформація, що може бути збережена у вигляді файлу зображення;
- восилання на сторінку, що містить візуальну або текстову інформацію;
- інформація міститься у файлі зручному для подальшого використання форматі, посилання на скачування якого представлено на ресурсі.

Також найбільш важливим критерієм для аналізі ресурсів є їх відкритість для користувача. Саме відкритість системи визначає не тільки час перебування на ресурсі користувача, досягнення кінцевої мети пошуку, а також повернення до ресурсу для здійснення нових пошуків.

З точки зору відкритості системи було визначено наступні характеристики, що визначають доступність інформації користувачеві:

а) ресурс відкритого доступу, дані на якому повністю відкриті для перегляду та завантаження:

- доступ до ресурсу не передбачає реєстрації користувача;
- доступ до ресурсу передбачає реєстрацію користувача.

б) на ресурсі представлена частково інформація про наукові дані відповідно до переліку компонентів, що характеризують публікацію у скороченому вигляді. В свою чергу на ресурсі представлено посилання на більш повний контент.

в) доступ до ресурсу є передплачений, інформація доступна тільки після сплати організацією чи користувачем внеску за користування системою. До цього – ресурс є закритим для перегляду.

Таким чином після формування переліку критеріїв запропонованих для проведення аналізу ресурсів наукової інформації стало можливим запропонувати чотирирівневу описову модель ресурсу наукової інформації (рисунок 2.1).

Перший рівень передбачає опис відкритості системи. Представляє собою змінну деяку OA, що може набувати значення «0» чи «1», відповідно, якщо система відкритого доступу змінна набуває значення «1» і навпаки.

Другий рівень – множина набору даних – DS, що містить варіанти наборів даних:

$$DS = \{ds1, ds2, ds3, \ldots, dsn\}.$$

Чим більше наборів даних, тим більше інформації представлено на ресурсі, і ресурс в свою чергу задовольняє запити користувача щодо наукової інформації.

Пошук інформації для аналізу ресурсів цікавив з точки зору присутності рядка для пошуку на головній сторінці ресурсу. Тобто значення S може набувати значення «0» або «1».

І четвертий рівень – збереження інформації. В даній моделі передбачено, що інформація, яка може бути збережена – це або повний текст документу або реферативна інформація про публікацію.

$$ID = \{FI, AI\},$$

де FI – можливість завантаження повнотекстового документу, AI – можливість завантаження реферативної інформації.

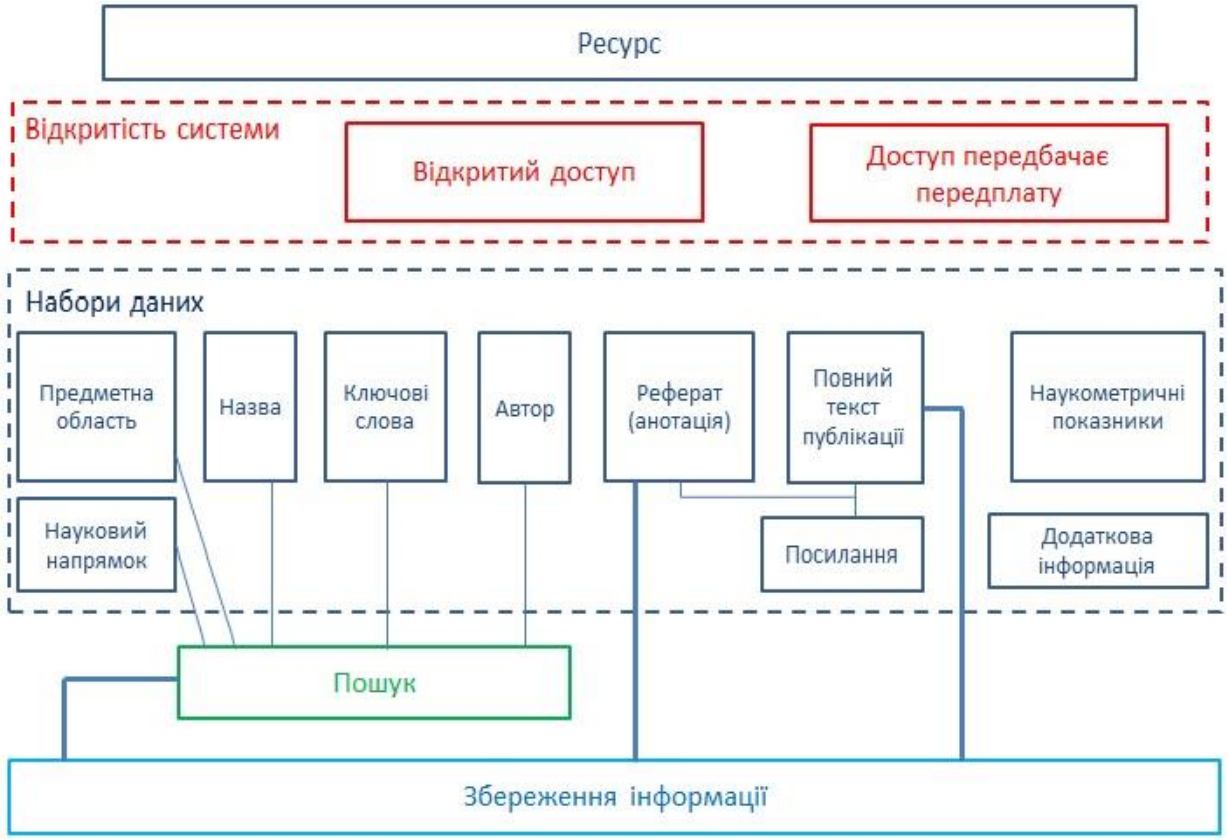


Рисунок 2.1 – Чотирерівнева описова модель ресурсу наукової інформації

Аналіз систем наукової інформації, що були обрані експертним шляхом було проведено за усіма запропонованими параметрами.

Наукометричні ресурси.

Web of Science (компанія Clarivate Analytics) – є сукупністю баз даних, що представляють інформацію про публікації більш ніж 13 тисяч авторитетних наукових видань та патентів на винаходи. Крім інформації про публікації та патенти Web of Science є джерелом великого масиву наукометричних даних, що обраховуються для публікацій та патентів – імпакт фактор журналу, цитованість, індекс Гірша тощо.

Згідно критеріїв, які було запропоновано для оцінки ресурсів наукової інформації можна зробити висновок щодо системи: система є передплаченою і не передбачає доступ до системи всіх користувачів мережі Інтернет, сторінка ресурсу є абсолютно закрита для перегляду користувачів, що не мають передплатного

доступу [55], (рисунок 2.2). Для даної системи показник OA набуває значення «0», OA=”0”.

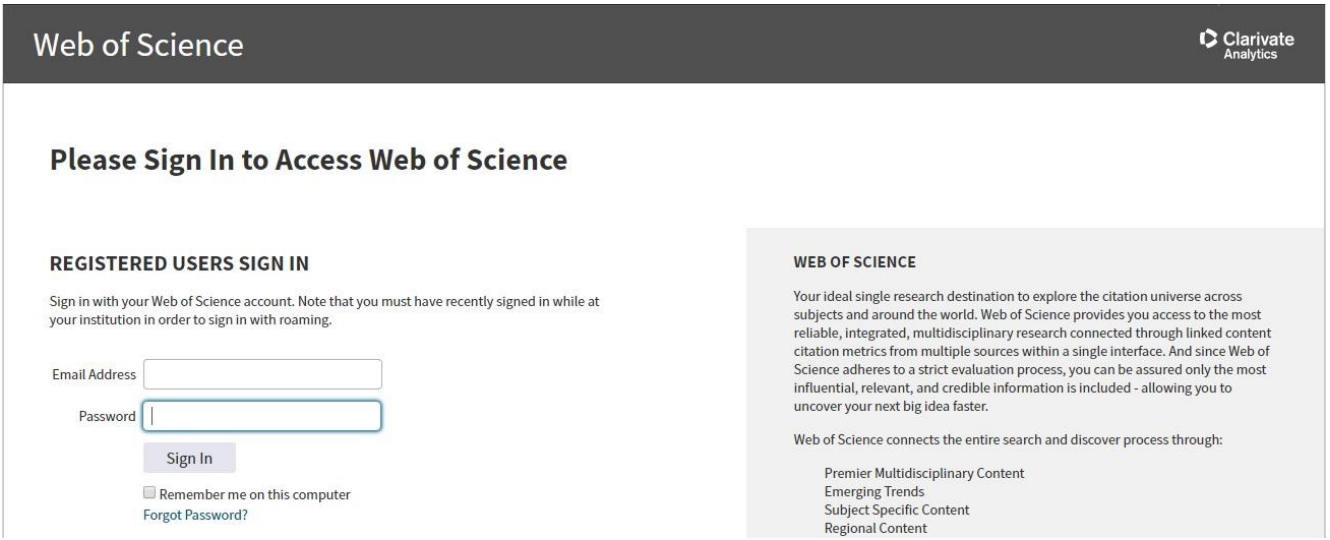


Рисунок 2.2 – Головна сторінка ресурсу Web Of Science, що відображається пересічному користувачеві мережі Інтренет

При наявності передплати для доступу до ресурсу (передплата може бути лише інститутційною) доступ до інформації можливий без додаткової реєстрації, в той же час кожен користувач може створити власний профіль на ресурсі для збереження історії своїх пошуків.

Серед наборів даних, що присутні на ресурсі відповідно до запропонованої моделі - предметна область, науковий напрямок, назва публікації, ключові слова, автор, анотація, наукометричні показники. В той же час представлення повного тексту документів ресурс не передбачає, в той же час можливий перехід за посиланням на сайт видавництва видання, чи інший ресурс, що надає більше інформації.

Щодо пошуку параметр S = «1», пошук інформації запроваджено на головній сторінці ресурсу.

Система передбачає проведення вибірки та збереження обраної інформації, або направлення обраних для збереження даних на електронну пошту користувача.

Варто зазначити, що система не передбачає API (Application Programing Interface) і тому розробка надбудов для отримання інформації є неможливим.

Scopus (компанія Elsevier) - є складовою інтегрованого науково-інформаційного середовища SciVerse. Системою проіндексовано більш ніж 22 тисячі наукових видань.

Відповідно до запропонованих критеріїв аналізу ресурсів можна зазначити наступне: доступ до даних системи передбачає передплату, тобто для даної системи показник ОА набуває значення «0», ОА=”0”. Але на головні сторінці для всіх користувачів Інтернет передбачений режим ознайомленні із інформацією, яку можна отримати на ресурсі, та доступ до певного набору метрик (рисунок 2.3) [56].

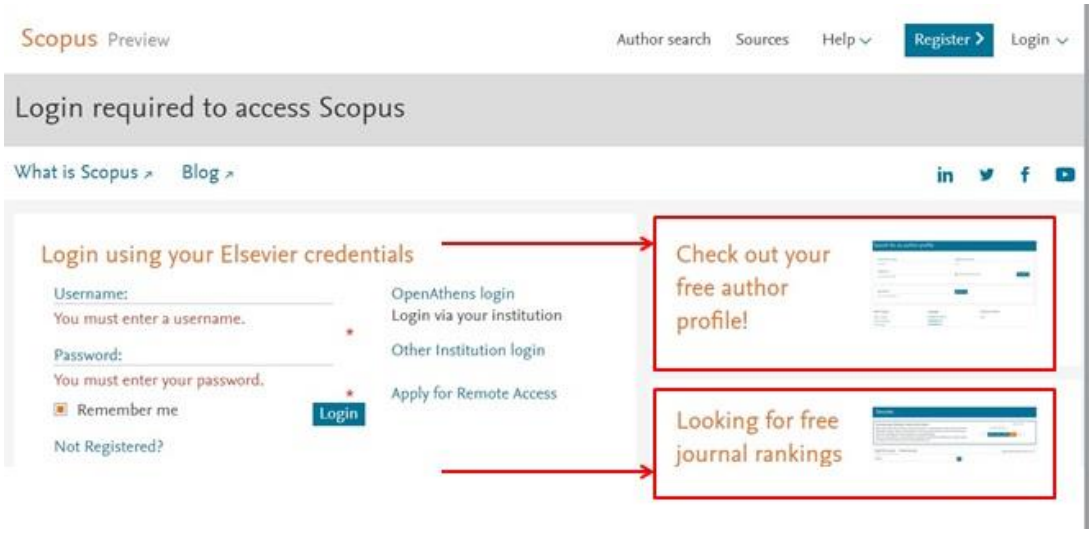


Рисунок 2.3 – Интерфейс головної сторінки ресурсу Scopus, що відображається пересічному користувачеві мережі Інтернет

Варто зазначити, що одним із безкоштовних ресурсів доступу до наукової інформації компанії Elsevier є ресурс Science Direct, що дозволяє здійснювати пошук публікацій за заданими параметрами. В той же час для зареєстрованих користувачів пропонує сервіс пропозиції рекомендованих публікацій відповідно до попередніх пошуків зареєстрованого на ресурсі користувача (рисунок 2.4) [57].

Структура та реалізація пошуку на вищезазначених ресурсах є подібна до іншних, зокрема Google Scholar, що надає можливість швидкого доступу до інформації, що знаходиться у відкритому доступі та не передбачає передплати [58].

Серед наборів даних доступних користувачеві на ресурсі Scopus є наступні: предметна область, науковий напрямок, назва публікації, назва видання, автор, ключові слова, анотація. Повних текстів документів на ресурсі не передбачено.

Пошук інформації запроваджено на головній сторінці ресурсу, тобто параметр, що визначає цей критерій аналізу - S = «1». Але це відноситься до користувачів мережі Інтернет, що мають передплачений доступ до ресурсу.

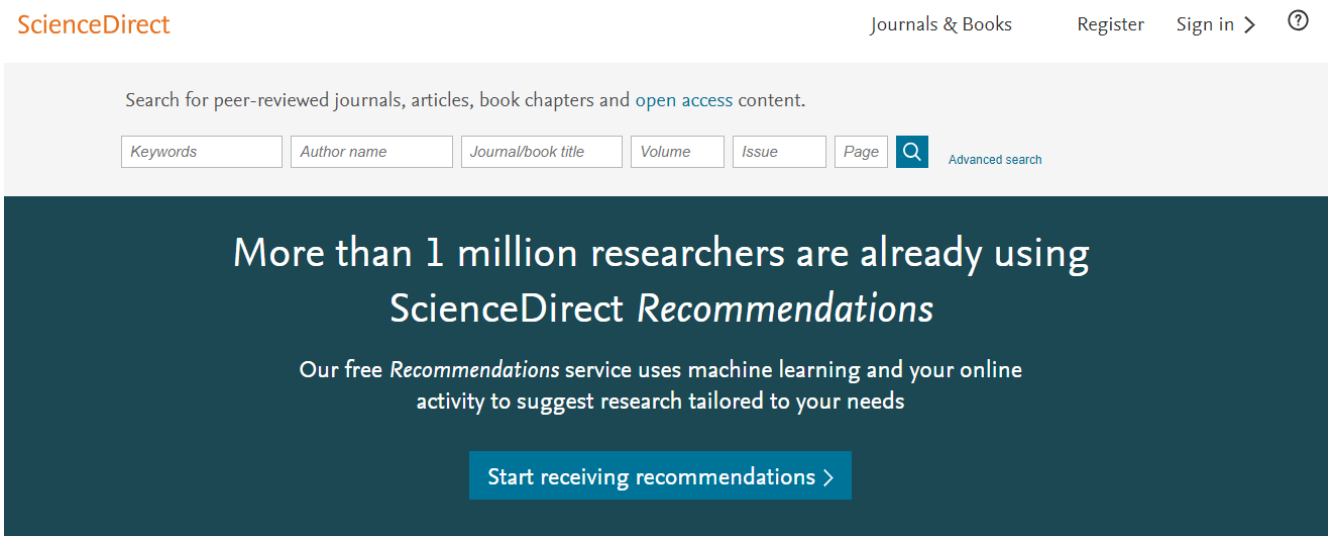


Рисунок 2.4 – Головна сторінка ресурсу Science Direct

Система передбачає можливість збереженні інформації, що обрана користувачем для подальшої роботи.

Система не дозволяє розробку додаткових програмних засобів для формування масивів інформації на базі ресурсу.

Google Scholar (компанія Google) – це не тільки сервіс пошуку наукової інформації в мережі Інтернет (рисунок 2.5) [59], а також система, що індексує повнотекстові метадані наукової літератури та представляє ряд метрик – зокрема індекс Гірша, кількість цитувань, тощо.

Регулярно ресурс надає користувачам мережі Інтернет інформацію про рейтинг публікацій за кількістю цитувань з наступних наукових напрямків (рисунок 2.6) [60]:

- Харчові технології;
- Джерела енергії;

- Інженерія та комп’ютерні науки;
- Гуманітарні науки, література та мистецтво.

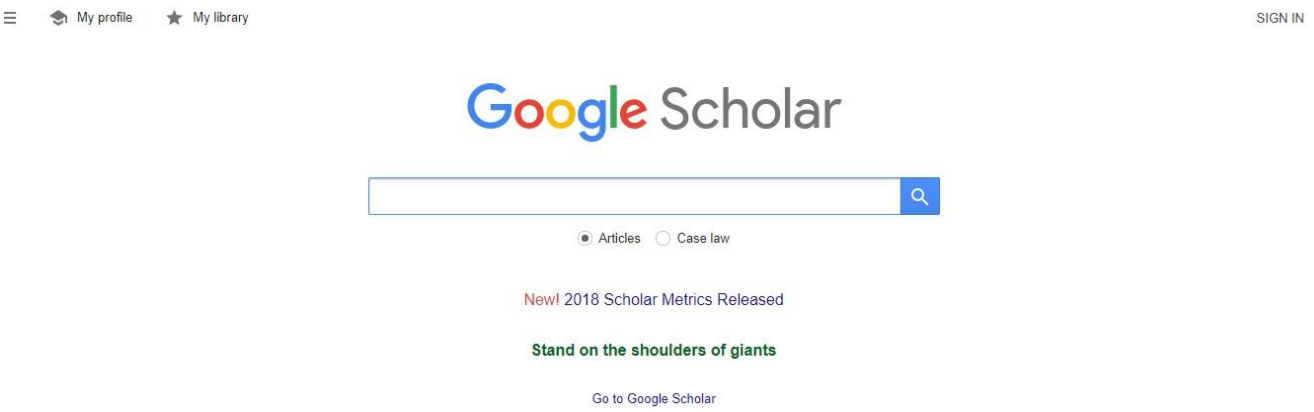
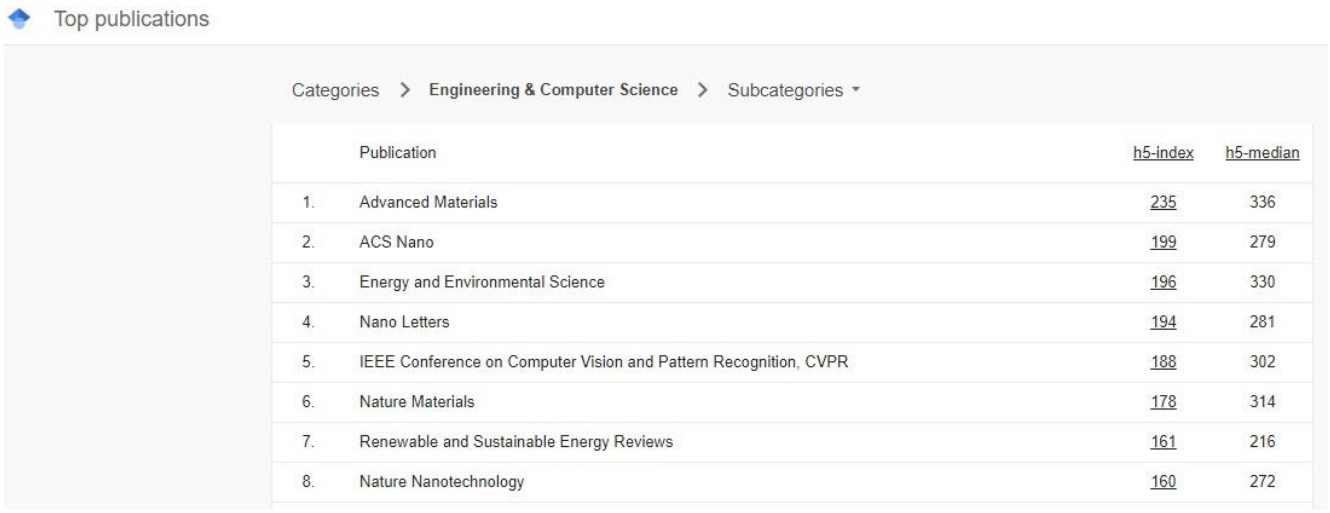


Рисунок 2.5 – Головна сторінка ресурсу Google Scholar

Сервіс абсолютно відкритий для доступу до інформації. Кожен користувач може створити на ресурсі Google Scholar Citations власний профіль і відслідковувати кількість цитувань та наукометричні показники власного публікаційного доробку. Приклад профілю автора Google Scholar Citations зображено на рисунку 2.7 [61].



	Publication	h5-index	h5-median
1.	Advanced Materials	235	336
2.	ACS Nano	199	279
3.	Energy and Environmental Science	196	330
4.	Nano Letters	194	281
5.	IEEE Conference on Computer Vision and Pattern Recognition, CVPR	188	302
6.	Nature Materials	178	314
7.	Renewable and Sustainable Energy Reviews	161	216
8.	Nature Nanotechnology	160	272

Рисунок 2.6 – Перелік видань із найбільшим 5-річним індексом Гірша за напрямком Інженерія та комп’ютерні науки.

На сторінці автора відображається перелік публікацій, який можна оновлювати вручну, а також використовувати автоматичний пошук публікацій в системі. Також відстежувати загальну кількість цитувань всіх публікацій, індекс Гірша автора, та 10-річний індекс Гірша. Також система надає інформацію про

кількість цитувань кожної публікації зазначеної в переліку. Система надає можливість отримувати на електронну пошту оновлену інформацію свого профілю – нові цитування, тощо.

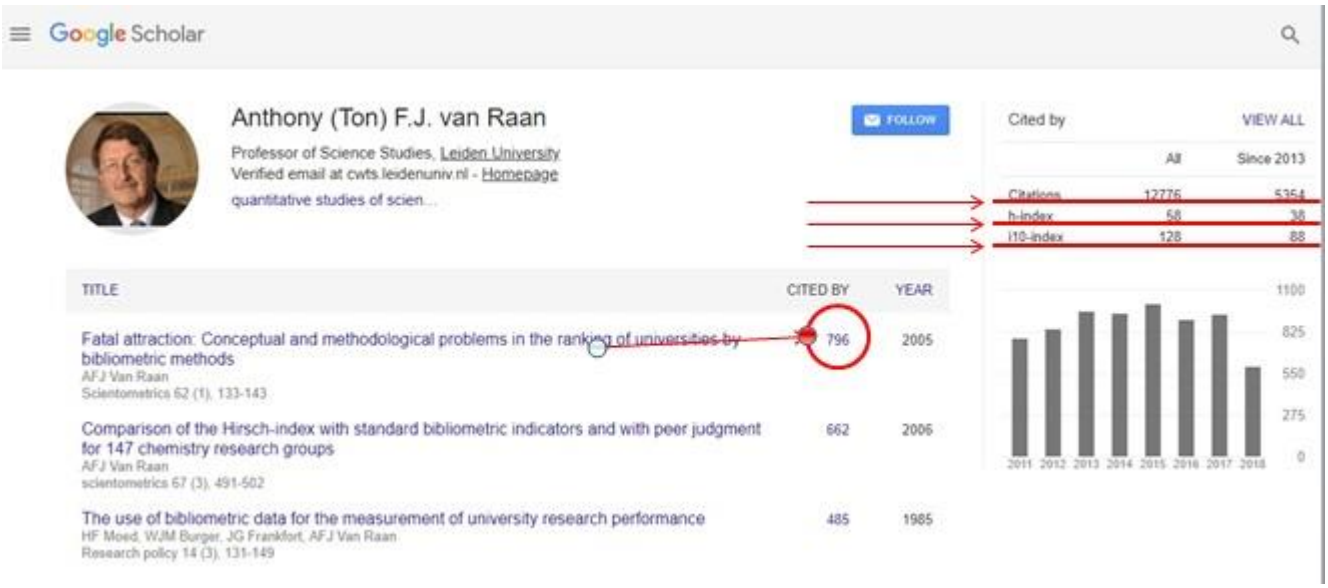


Рисунок 2.7 – Приклад сторінки автора в системі Google Scholar Citations

Значною перевагою сервісу є швидке оновлення інформації про публікації, які індексуються системою.

Значну увагу дослідників, що вивчають наукометричні параметри та системи – провайдери наукометричної інформації порівнюють Google Scholar з такими ресурсами як PubMed, Web Of Science, Scopus – недоліки та переваги [12], особливості систем Web Of Science, Scopus та Google Scholar, який саме індекс Гірша брати до уваги [14].

Сервіс привертає увагу з точки зору інформації, яку користувач може зазначати при створенні, редагуванні та опрацюванні власного профілю, зокрема – сфера наукових інтересів, афіліація, електронна пошта, а також – зазначати співавторів. На сьогодні пошук співавторів є задачею, що цікавить науковців з точки зору побудови нових колаборацій, то міждисциплінарних дотичних для реалізації глобальних проектів. Багато доробків присвячено безпосередньо побудові мереж співавторів для подальшої картини співпраці науковців різних галузей знань, а також в межах вузьких наукових напрямків [62, 63, 64].

Відповідно до критеріїв запропонованих для аналізу систем наукової інформації Google Scholar є системою абсолютно відкритою, що надає величезні масиви проіндексованої наукової інформації і дозволяє кожному користувачеві створити власний профіль для опрацювання масиву власного публікаційного доробку, і в той же час знайти інформацію про публікаційну активність науковців, за напрямками, що становлять сферу інтересів.

Серед наборів даних, що репрезентує система наступні – предметна область, науковий напрямок, автор, назва публікації, назва та атрибути видання.

Повного тексту документу на ресурсі на представлено, але в той же час посилання на повний текст документу – присутнє.

Також представлені наукометричні показники. Пошук запроваджено на головній сторінці ресурсу.

Можливості збереження обраної в результаті пошуку інформації немає. Збереження можливе тільки за рахунок реалізації screen-shot зображень.

Інтерфейс є зручним для користування і, не дивлячись на обмеженість уточнення пошуку – результати зображуються у зручному для перегляду вигляді (рисунок 2.8) [65] із відповідними гіперпосиланнями на сторінки автора і короткими наукометричним даними щодо цитувань публікації та переходом до повного тексу публікації або реферованої інформації на сторінці видавця.

Також можна із впевненістю зауважити, що система Google Scholar Citations представляє інтерес для розробника – є джерелом даних, що можуть бути опрацьовані і на їх базі сформовані та візуалізовані нові масиви інформації, що можуть являти собою корисні інструменти, та оптимізувати роботу із системою в пошуках даних.

На сьогоднішній день переважна більшість світових університетів та наукових організацій створюють та активно використовують інститутційні репозитарії. Також серед інформаційних систем наукової інформації в мережі інтернет розповсюджені репозитарії публікацій з вузьких галузей знань.

При проведенні аналізу систем наукової інформації було розглянуто настпуні репозитарії.

CSIRO's Research Publications Repository, JRC Publications Repository, Research and Publications Archive (The David and Yoalnda Katz faculty of arts).

CSIRO - The Commonwealth Scientific and Industrial Research Organization – Спілка промислових та дослідницьких організацій – є національною агенцією Австралії та однією з найбільших дослідницьких агенцій у світі.

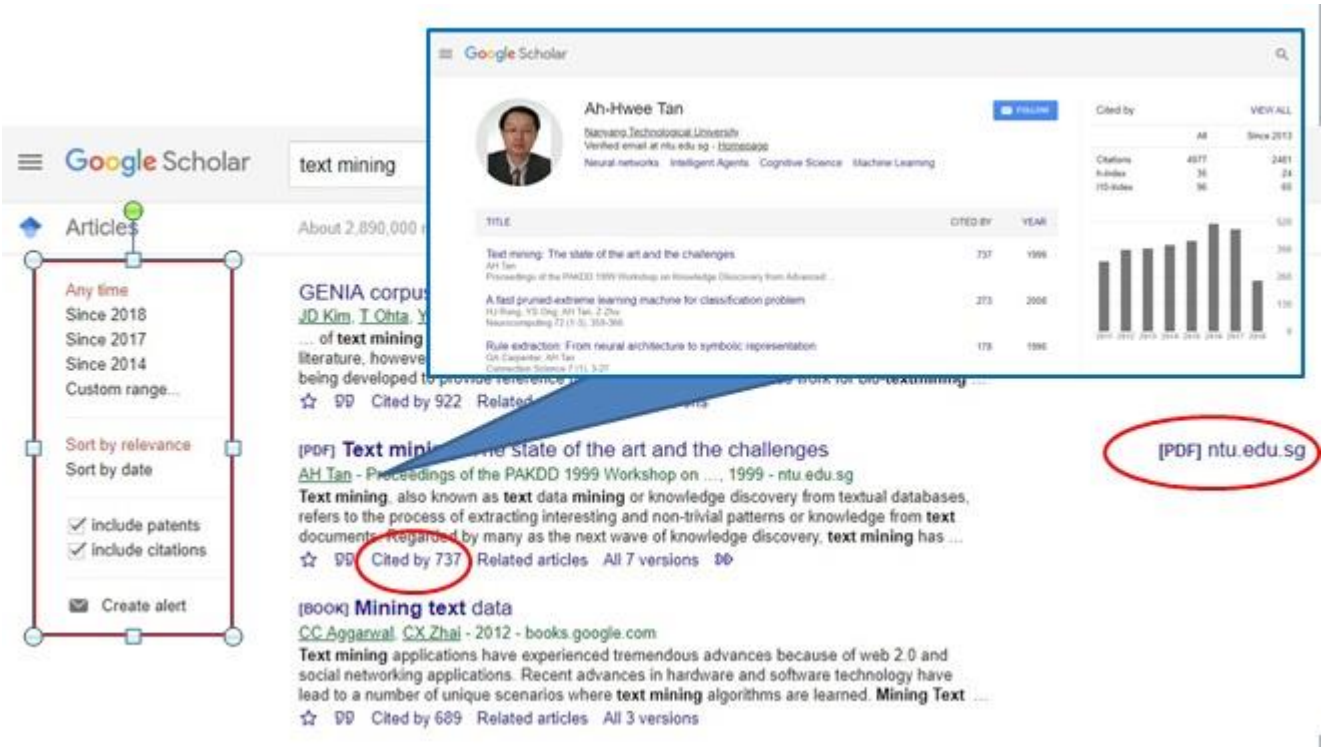


Рисунок 2.8 – Приклад сторінки – результатів пошуку за заданим концептом ресурсу Google Scholar

Ресурс є відкритого доступу, реєстрація на ресурсі для доступу до інформації непотрібна. Пошук інформації організований на головній сторінці ресурсу (рисунок 2.9) [66].

Доступ за логіном і паролем до системи можливий тільки за рахунок реєстрації, але для окремих користувачів системи – за допомогою служби підтримки ресурсу.

Серед наборів даних, що представлені в репозитарії за результатами пошуку відображаються наступні:

- назва публікації;
- автор;

- дата публікації;
- тип публікації;
- назва журналу та ідентифікатори;
- анотація;
- ключові слова;
- мова публікації.

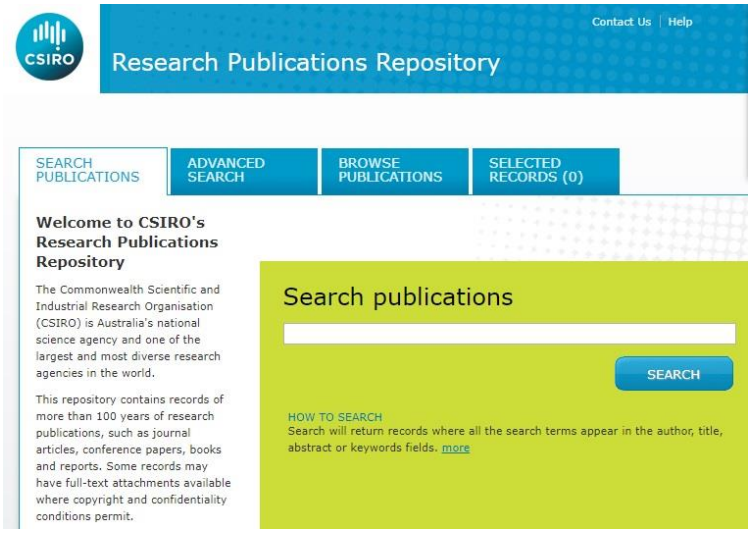


Рисунок 2.9 – Головна сторінка ресурсу CSIRO’s Research Publications Repository

На ресурсі також доступні показники цитувань публікації (рисунок 2.10) [66], на відміну від інших ресурсів серед показників цитованості є показники Altmetric [67] – показники цитованості публікації в соціальних мережах та альтернативних науковим ресурсах.

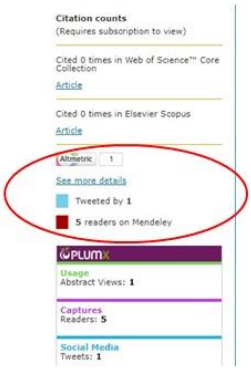


Рисунок 2.10 – Показники цитованості публікації в наукометричних ресурсах та альтернативних інтренет-системах.

Опція збереження тексту публікації не передбачено.

Єдиний дослідницький центр (Joint Research Centre) – це знаннєво-науковий сервіс Європейської Комісії. В свою чергу JRC Publications Repository – он-лайн сервіс, що настає доступ до даних про наукові публікації Єдиного дослідницького центру Європейської комісії.

Варто відмітити, що JRC Publications Repository є частиною офіційного порталу Європейської Комісії.

Відповідно до політик Європейської Комісії щодо публікування наукових результатів – ресурс є відкритим для всіх користувачів Інтернет та не передбачає реєстрації для отримання більшої інформації чи розширених можливостей для користувачів системою.

На головні сторінці передбачено три варіанти пошуку для користувача (рисунок 2.11) [68]:

Рисунок 2.11 – Головна сторінка репозитарію Єдиного дослідницького центру із опціями пошуку для користувача.

- Звичний для інтерфейсу веб-сторінок в правому верхньому куті сторінки;
- Додатковий рядок із підказками щодо введення інформації для пошуку;
- Посилання для пошуку за розділами та додатковим розгорнутим пошуком та уточненням для пошуку по розділах (рисунок 2.12) [68].

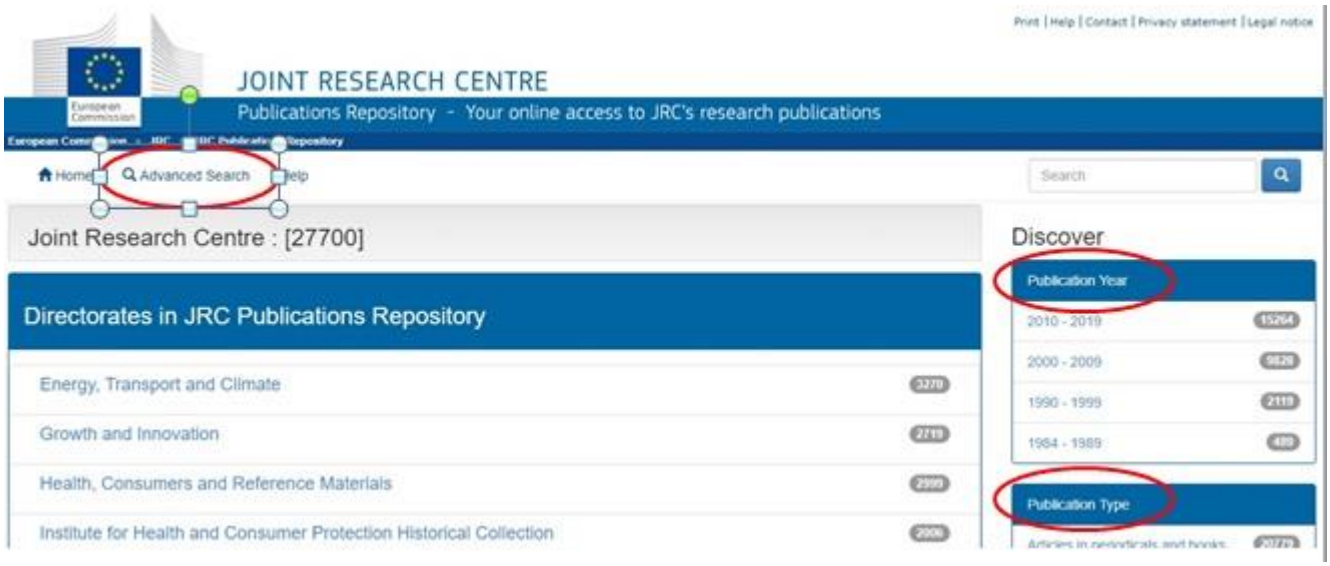


Рисунок 2.12 – Сторінка пошуку публікацій Єдиного дослідницького центру із опціями уточнення для пошуку та розгорнутого пошуку.

Ресурс передбачає надання даних про публікації за наступними напрямками:

- Energy, Transport and Climate
- Growth and Innovation
- Health, Consumers and Reference Materials
- Institute for Health and Consumer Protection Historical Collection
- Joint Research Centre Corporate Activities
- Joint Research Centre Historical Collection
- Nuclear Safety and Security
- Space, Security and Migration
- Sustainable Resources

Кожна публікація серед запропонованих до аналізу ресурсів наукової інформації присутні наступні: науковий напрямок, автор, назва видання та його реквізити, анотація. Наукометричних показників, повного тексу публікації або посилання на нього на ресурсі не передбачено, репозитарій передбачає виключно інформацію про публікацію.

Research and Publications Archive (The David and Yoalnda Katz faculty of arts) – один із небагатьох ресурсів наукової інформації, що присвячений. Безпосередньо

архів є частиною ресурсу факультету мистецтв Девіда та Йоланди Катц університету Тель-Авіва (рисунок 2.13) [69].

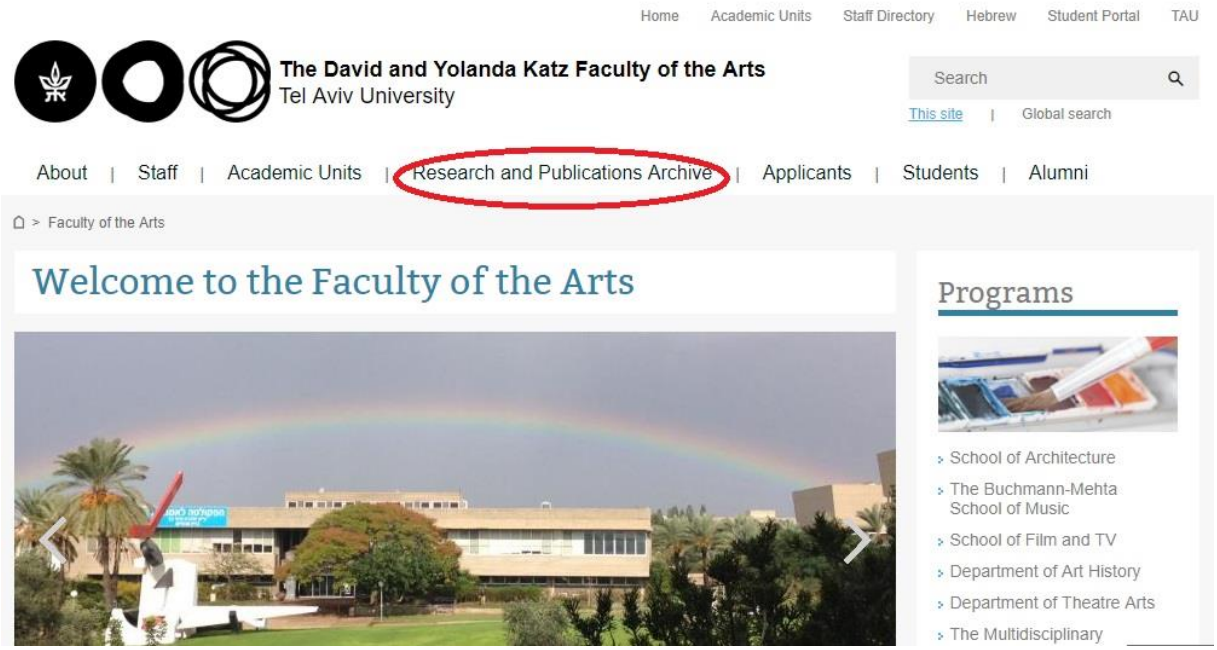


Рисунок 2.13 – Головна сторінка The David and Yoalnda Katz faculty of arts

Ресурс не передбачає уточнення пошуку чи структуровану інформацію про публікацію. Інформація є загальною, містить дані про автора, назву видання і назву публікації. Завантаження інформації не передбачено.

arXiv – самий популярний архів публікацій в мережі Інтернет, адже дозволяє науковцям розміщувати препринти своїх публікацій із опцією подальшої зміни характеристик документу відповідно до того, де публікацію було розміщено в подальшому, або розміщувати публікації, що вже було надруковано у наукових виданнях.

Ресурс є абсолютно відкритим для перегляду та доступу до інформації. Створення власного профілю – реєстрація передбачена лише для розміщення документів на ресурсі.

Архів препринтів передбачає розміщення документів за наступними предметними областями:

- Physics
- Mathematics
- Computer Science

- Quantitative Biology
- Quantitative Finance
- Statistics
- Electrical Engineering and Systems Science
- Economics

Кожна предметна область передбачає ряд наукових напрямків, за якими можна уточнити пошук чи розмістити публікацію.

Пошук розміщено на головній сторінці ресурсу. Передбачає пошук по всіх розділах та по визначеному користувачем розділу (рисунок 2.14) [70].

Важливо зауважити, що інтерфейс ресурсу регулярно (приблизно один раз на кілька місяців) оновлюється.

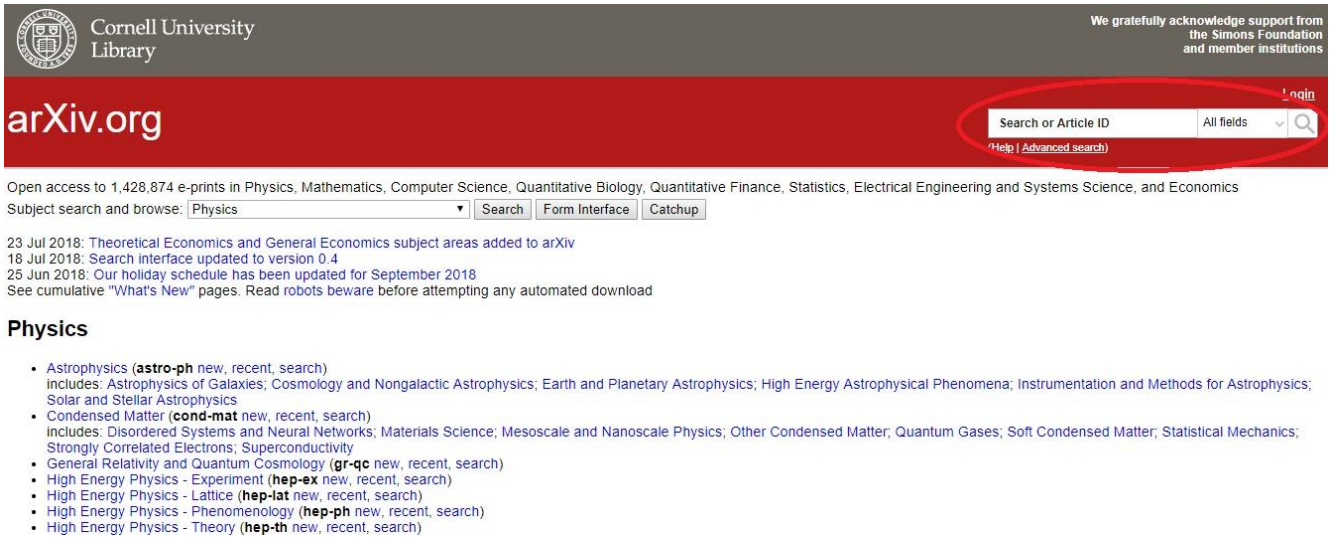


Рисунок 2.14 – Головна сторінка архіву препринтів arXiv

Серед наборів даних, що становлять інтерес для користувача інформаційними ресурсами наукової інформації, архів передбачає наступні: предметна область, науковий напрямок, назва публікації, автор, анотація, інформація про видання у випадку, якщо документ опубліковано у науковому виданні.

Жодної наукометричної інформації чи показників ресурс не надає. В той де час можливе завантаження повного тексту документу у форматі .pdf, але можливості зберегти результати пошуку ресурс не передбачає.

Ресурс є відкритим для розробників і за допомогою HTML developers window можливе створення нових додатків для обробки інформації ресурсу та створення нових інформаційних масивів та їх подальшої візуалізації.

Zenodo – ресурс, що містить відкриті дані, публікації, презентації та коди програм проектів.

Пошук та завантаження продуктів на ресурс здійснюється на головній сторінці системи. Користувач може бути створити власний профіль в системі, проте, можливий вхід до системи із логіном на паролем ресурсу відкритих програмних кодів – GitHub (рисунок 2.15) [71].

Корисутвачеві, що здійснює пошук інформації за результатами пошуку стає доступним наступний набір інформації про наукову публікацію – назва, автор(и), коротка анотація та посилання для завантаження повного тексту документу. Також у окремій частині сторінці зазначається дата публікації, DOI, ключові слова та назва журналу, в якому було розміщено роботу.

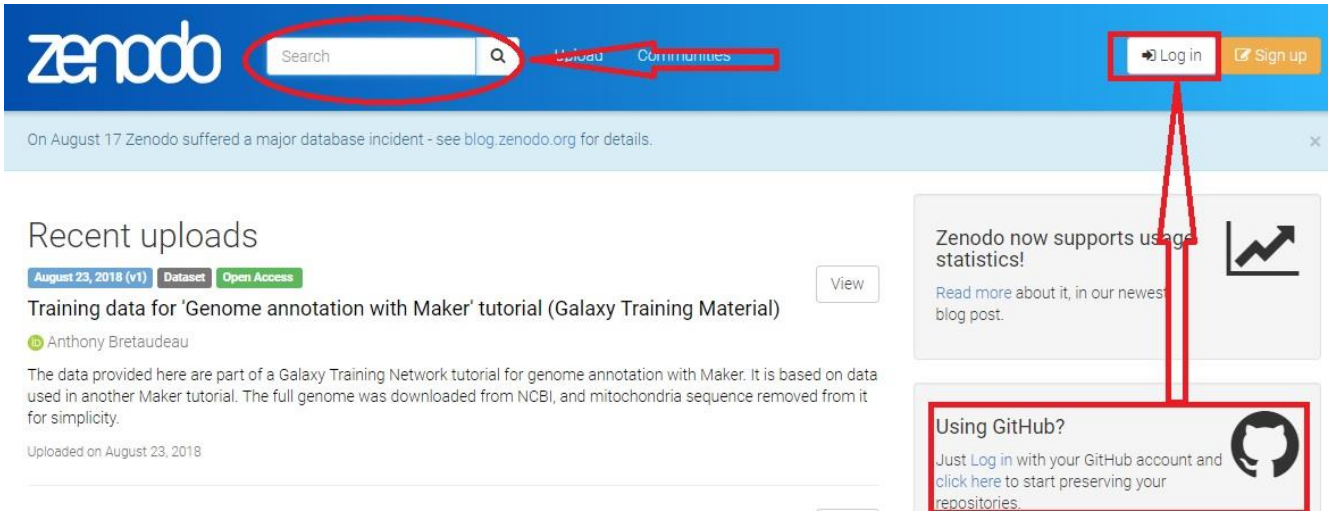


Рисунок 2.15 – Головне сторінка ресурсу Zenodo із позначенням рядка пошуку та можливості входу із персональними даними ресурсу GitHub.

Система передбачає обрахунок лише кількості переглядів та завантажень публікації, жодних наукометричних показників на ресурсі не представлено.

Важливо звернути увагу, що ресурс є абсолютно відкритим і, навіть, інформація про систему зазначає відкритість коду, що в свою чергу робить можливим розробку та застосування додатків для отримання додаткових масивів інформації.

Springer – один із провідних світових видавців наукової літератури. Видавництво видає наукову літературу з усіх галузей знань – це 2,900 журналів та 290,000 книг. Ресурс є відкритим для перегляду, але можливо пройти процедуру реєстрації на ресурсі для отримання повідомлень та новин щодо інформації, яка цікавить користувача і в той же час – отримувати інформацію про акційні пропозиції на наукову літературу.

Пошуковий рядок розміщений в традиційній для веб-сторінок локацій – правий верхній кут, але уточнення пошуку не є достатнім для пошуку наборів документів за заданим концептом (рисунок 2.16) [72]. Тобто пошук інформації на ресурсі здійснюється шляхом поступового переходу за гіперпосиланнями.

Серед даних, що цікавлять користувача на ресурсі видавництва Springer наступні набори даних: назва публікації, інформація про видання та відповідні реквізити, анотація і ключові слова. Серед метрик для публікації – лише кількість звернень до публікації.

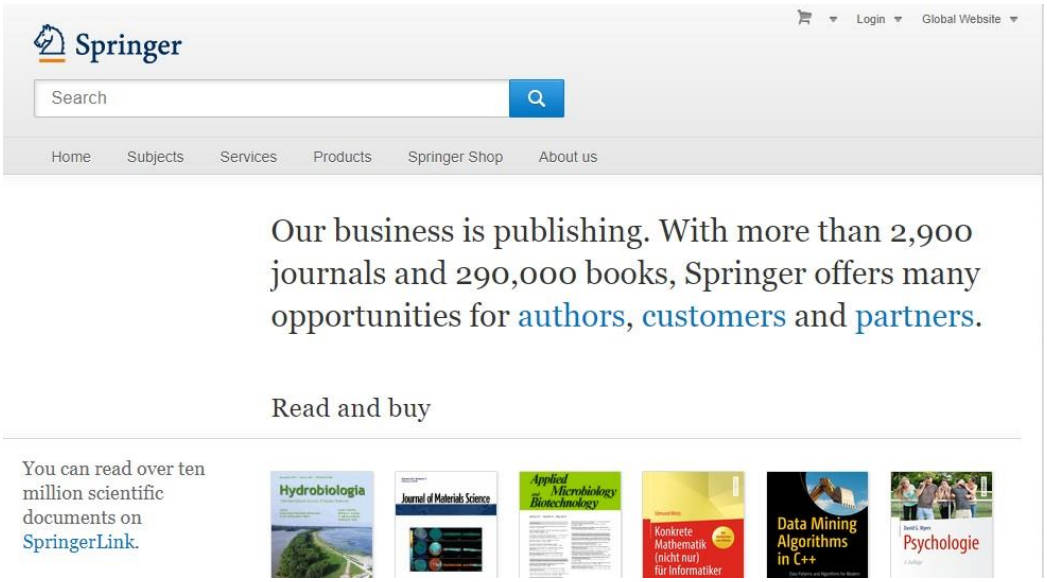


Рисунок 2.16 – Головна сторінка ресурсу видавничої літератури Springer

В той же час наукометричні показники відповідно до ресурсів, що індексують наукові видання наявні для кожного наукового видання представленого на ресурсі.

Варто зауважити, що серед видань видавництва є наукова література відкритого доступу, відповідно значна частина публікації, що надає уявлення про матеріал розміщено у вікні, що представляє інформацію про публікацію, а також містить посилання на завантаження повного тексту публікації.

Elsevier – поряд із Springer є одним із провідних світових видавців наукової літератури.

Пошук безпосередньо журналів, книг та веб-сторінок пов’язаних із заданим концептом розміщено на головній сторінці ресурсу, в той же час пошук публікацій здійснюється за допомогою сервісу компанії Elsevier – ScienceDirect (рисунок 2.17) [73].

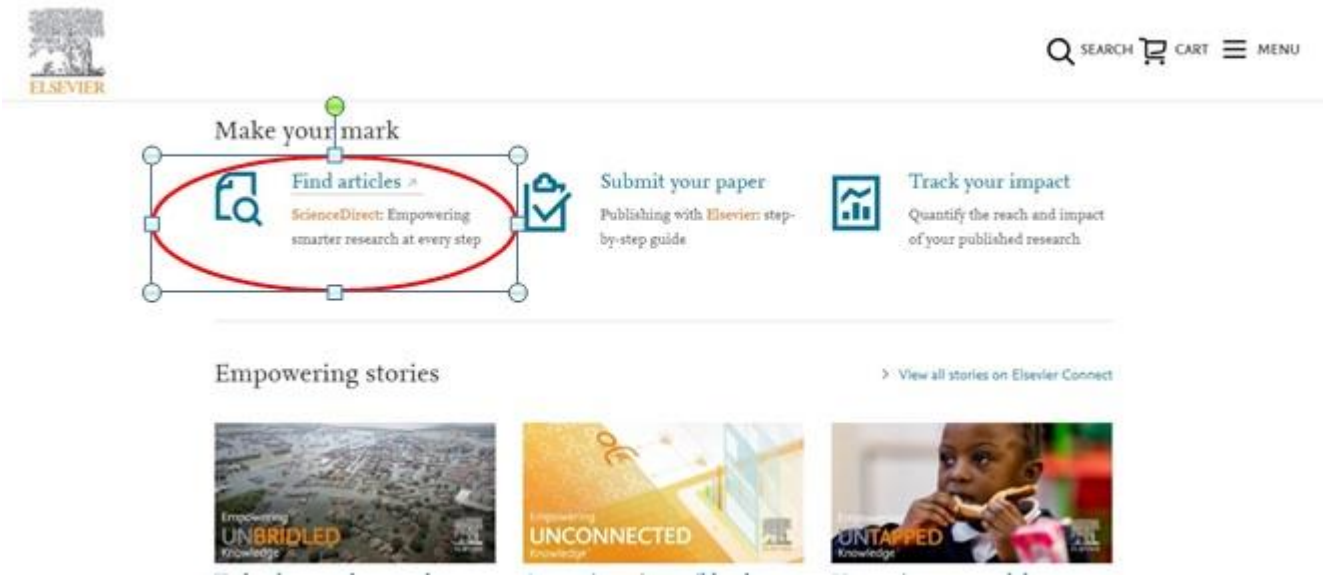


Рисунок 2.17 – Головна сторінка видавництва Elsevier – посилання на пошук публікацій на ресурсі ScienceDirect

Ресурс є абсолютно відкритим для користувачів. Реєстрація синхронізується із іншими ресурсами компанії Elsevier, зокрема Scopus.

При пошуку публікацій система розділяє результати пошуку по предметних областях, що з одного боку – структурує пошук, в той же час надає користувачеві можливість користувачеві оцінити приналежність концепту заданого для пошуку іншим предметним областям.

Всі публікації, що знаходяться у відкритому доступі відкриті для завантаження у форматах *.pdf та *.doc. Також присутні опції збереження тільки реферативної інформації про публікацію або переліку посилань та експортувати їх у інші ресурси – Mendeley або Refworks у запропонованих форматах (рисунок 2.18).

Система дозволяє користувачеві отримати повний набор даних, що може цікавити користувача: назва публікації, видання та його реквізити, ключові слова, анотації, повний текст публікації.

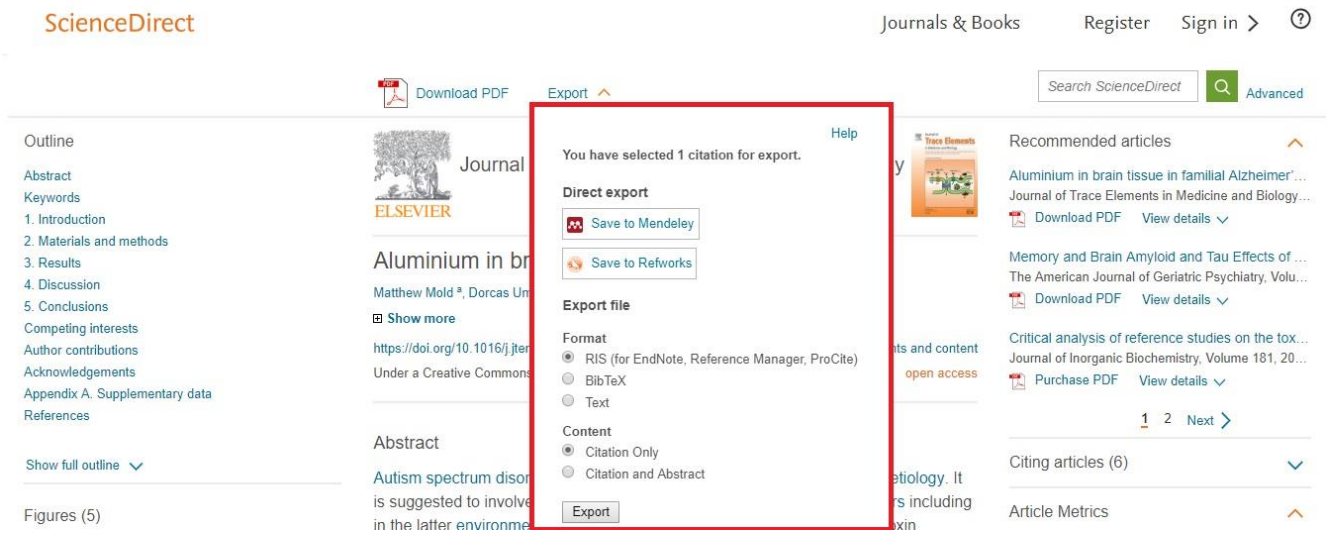


Рисунок 2.18 – Опції для збереження інформації отриманої в результаті пошуку в рамках ресурсу Science Direct

Серед наукометричних показників, що можуть зацікавити користувача – індекс цитування та кількість згадувань публікації у соціальних медіа та публікації, що посилалися надану роботу.

International Organization of Scientific Research – Міжнародна організація наукових досліджень здійснює підтримку освітянам та дослідникам по всьому світу, зокрема країнам, що розвиваються. Це є асоціація науковців, дослідників, інженерів, менеджерів з у сього світу. Організація є видавцем журналів відкритого доступу з різних галузей знань. Відповідно до політик організації – виконує індексацію наукових видань – для кожного з них обраховується імпакт-фактор.

Ресурс є відкритим для всіх користувачів мережі Інтернет. Реєстрацію користувачів в системі не передбачено.

Пошуковий сервіс на головній сторінці не передбачено, в той же час на головній сторінці ресурсу представлено перелік видань, якими опікується організація із зазначенням імпакт-фактору журналу (рисунок 2.19) [74].

Сервіс надає можливість користувачеві отримати інформацію лише про видання, умови публікації та тематику. Не зважаючи на те, що ресурс анонсує

видання лише відритого доступу – перехід по посиланнях для пошуку інформації про публікації є край ускладненим.

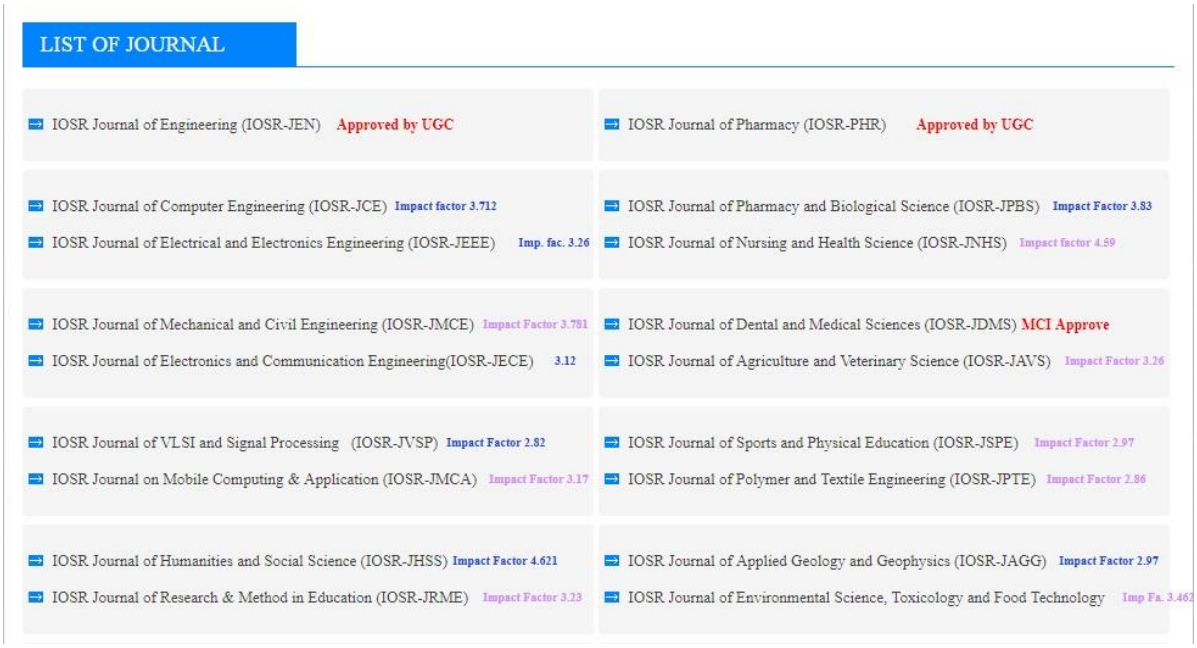


Рисунок 2.19 – Перелік журналів, що видає International Organization of Scientific Research із зазначенням імпакт-фактору.

Серед наукометричних показників, що представлені в системі можна зазначити лише імпакт-фактор журналів, що видає організація.

Scientific Research – видавництво наукової літератури відкритого доступу. Ресурс є відкритим для перегляду інформації всім користувачам мережі Інтернет. На відміну від інших ресурсів, що було оглянуто до цього – на головній сторінці присутня опція переведення веб-сторінки в режим мобільних пристроїв.

Пошук запроваджено на головній сторінці системи із підказками щодо контенту, який може бути введений для пошуку інформації.

Серед наборів даних, що надає система, як результат пошуку наступні: назва публікації, видання та необхідні реквізити, ключові слова, анотація та перелік посилань, на які посилався автор у публікації

В зручний спосіб представлена можливість завантажити повний текст публікації – або у форматі .pdf або представлення повного тексту публікації на окремій веб-сторінці (рисунок 2.20) [75].

Для кожної статті зазначається кількість завантажень, також за відповідним посиланням – здійснюється перехід до сторінки, на якій відображається перелік посилань на дану публікацію, що формується за даними сервісу Google Scholar.

Рисунок 2.20 – Сторінка із зазначенням можливих опцій для збереження тексту публікації

Серед наукометричних показників для окремих наукових видань, що представлені на ресурсі – альтернативний імпакт-фактор, що обраховується із використанням даних – кількості цитувань – сервісу Google Scholar, 5-річний індекс Гірша, кількість цитувань, завантажень та переглядів.

З огляду на те, що ресурс розроблений та опублікований в мережі Інтернет за ліцензії 'Creative Commons – Attribute', що дозволяє створювати надбудови для системи, або додаткові програмні продукти для подальшої обробки, реструктуризації та візуалізації отриманої на ресурсі інформації.

Серед наукових видань для аналізу було обрано International Journal of Computer – наукове видання відкритого доступу, що передбачає публікацію рецензованих робіт з усіх напрямків Комп'ютерних систем. Сам ресурс є відповідно відкритим для усіх користувачів мережі Інтернет, в той же час на ресурсі передбачена реєстрація для користувачів, що мають набір подати публікацію до розгляду у видання.

Веб-ресурс видання є складним для користування. Не дивлячись на те, що пошуковий рядок розміщено на головній сторінці – він не привертає увагу користувача (рисунок 2.21) [76].

Рисунок 2.21 – Головна сторінка наукового видання International Journal of Computer із зазначенням сервісу пошуку для користувачів та імпакт-фактору журналу.

За результатами пошуку користувач може отримати наступні набори даних про публікацію: реквізити видання, в якому опублікований матеріал із пошуковим контентом, автор, назва публікації та посилання або на повний текст публікації або на анотацію. При переході до анотації або повного тексту стає доступними вся інформація про публікацію. Повний текст можливий для завантаження у форматі *.pdf.

В той же час деяка інформація має обмежений доступ і вимагає контакту із адміністратором для подальшого перегляду матеріалів.

Серед наукометричних показників лише імпакт-фактор журналу доступний користувачеві для ознайомлення на головній сторінці ресурсу.

З огляду на побудову веб-сторінки журналу розробка нових додатків для отримання нових масивів інформації або виокремлення наборів даних є ускладненим.

Open Science Journal [77] – визнане світовим науковим загалом онлайн наукове рецензоване мультидисциплінарне видання відкритого доступу. Головна сторінка ресурсу не передбачає пошуку, в той же час передбачає перехід до архіву номерів, де відповідно запроваджений пошук.

Система передбачає реєстрацію користувача для подання публікації до розгляду.

Серед наборів даних про публікацію користувачеві доступні: назва публікації, автор, анотація, ключові слова, перелік посилань, а також посилання на збереження повного тексту документу.

Наукометричні показники на ресурсі не передбачені.

Відповідно до ліцензії ‘Creative Commons – Attribute’ дані ресурсу можуть бути опрацьовані шляхом розробки додатків та проведення сканування ресурсу програмними засобами.

На сьогодні одними із актуальних ресурсів є ресурси не тільки наукових публікацій, але й наукових даних, що відповідно до політики надання грантової підтримки проектів мають бути розміщені у відповідних репозитаріях, про що зазначено на етапі подання запиту на грант.

OpenAire [78] – архів не тільки наукових публікацій, але й наукових даних і проектів, а також фондуючих організацій та провайдерів наукових даних, один із найбільш популярних серед науковців архівів для розміщення результатів своїх досліджень.

Ресурс є відкритим для доступу до нього користувачів мережі Інтернет, передбачає реєстрацію та створення власного профілю в системі.

Головне сторінка містить лише інформацію для користувачів мережі Інтернет про ресурс, передбачає пошук при переході за відповідним пунктом головного меню, або прогорнувши сторінку донизу, де також представлена опція розгорнутого пошуку.

За результатами пошуку користувачеві доступні наступні дані про публікацію: назва, автор, предметна область.

Важливо, що поряд із інформацією про публікацію за результатами пошуку стають доступні посилання на проект, в рамках якого було реалізовано дослідження, а також безпосередньо на наукові дані, що були отримані в результаті дослідження і стали базою для написання публікації.

Повний текст публікації можливий за посиланням представленим на ресурсі.

Наукометричні показники на ресурсі не представлені.

Варто відмітити, що як і попередні деякі вже описані ресурси OpenAire має ліцензію ‘Creative Commons – Attribute’, що дозволяє розробляти додатки та інший інструментарій для обробки даних, що представлені на ресурсі для створення нових масивів інформації щодо наукових даних, проектів та відповідних публікацій. Адже всі грантоотримувачі за програмами Європейської комісії зобов’язані розміщувати всі дані, що були отримані в рамках реалізації проекту у відкритому доступі. Саме дотримання політики відкритого доступу та відповідні ліцензії дозволяють розробляти надбудови для подальшої обробки відкритих даних.

В свою чергу семантично невід’ємною частиною ресурсів, що розміщують у відкритому доступі дані отримані за реалізації грантових проектів Європейської комісії є ресурси – архіви вихідних кодів програм, що були розроблені для реалізації грантових проектів.

Software Heritage [79] - один із нових ресурсів, що вміщує майже 5 мільйонів вихідних файлів та містить інформацію з:

- відкриті репозитарії з GitHub [80];
- вихідні пакети з дистрибутиву Debian [81];
- відкриті репозитарії зі служби хостінгу коду – Gitorious [82];
- відкриті репозитарії зі служби хостінгу проектів – Google Code [83];
- велізи з проекту GNU [84].

На сьогодні відкритість вихідних кодів є актуальним аспектом поряд із відкритими науковими даними. Адже кожен науковий проект передбачає не тільки реалізацію конкретної наукової задачі, а й вирішення проблеми автоматизованої складової.

CERN Scientific Information Service [85] – он-лайн архів Європейської організації з ядерних досліджень – міжнародний дослідницький центр в Європі із фізики високих енергій, є своєрідним пошуковим сервером для будь-якої інформації, що пов'язана із експериментами ЦЕРН та архівом он-лайн публікацій. Ресурс є відкритим для доступу користувачів мережі Інтернет.

Пошук інформації передбачений на головній сторінці ресурсу, а також передбачена можливість створення власного профілю або використовувати для входу в систему дані соціальних мереж, або поштових ресурсів.

За результатами пошуку є доступними: назва публікації/книги, автори, предметна область, анотація.

Інформація щодо наукометричних показників на ресурсі не передбачена.

Registry of Research Data Repositories – ресурс відкритого доступу, що передбачає пошук наукової інформації у більш ніж у 2000 репозитарієв Європи. Інтерфейс системи передбачає пошук на головній сторінці без представлення додаткової інформації для зручності роботи користувача із пошуковою системою.

Реєстрація користувачів на ресурсі не передбачена.

На відміну від інших пошукових систем, Registry of Research Data Repositories надає за результатами пошуку перелік репозитарієв та архівів, що містять дані за заданим ключовим словом чи словосполученням.

Окрім назви репозитарію чи архіву, користувач отримує наступну інформацію: предметна область, типи даних, що містить система, країна, коротка інформація про ресурс, а також графічну інформацію щодо відкритості архіву чи репозитарію, ліцензії, наявність персональних політик, а також дотримання певних стандартів (рисунок 2.22) [86].

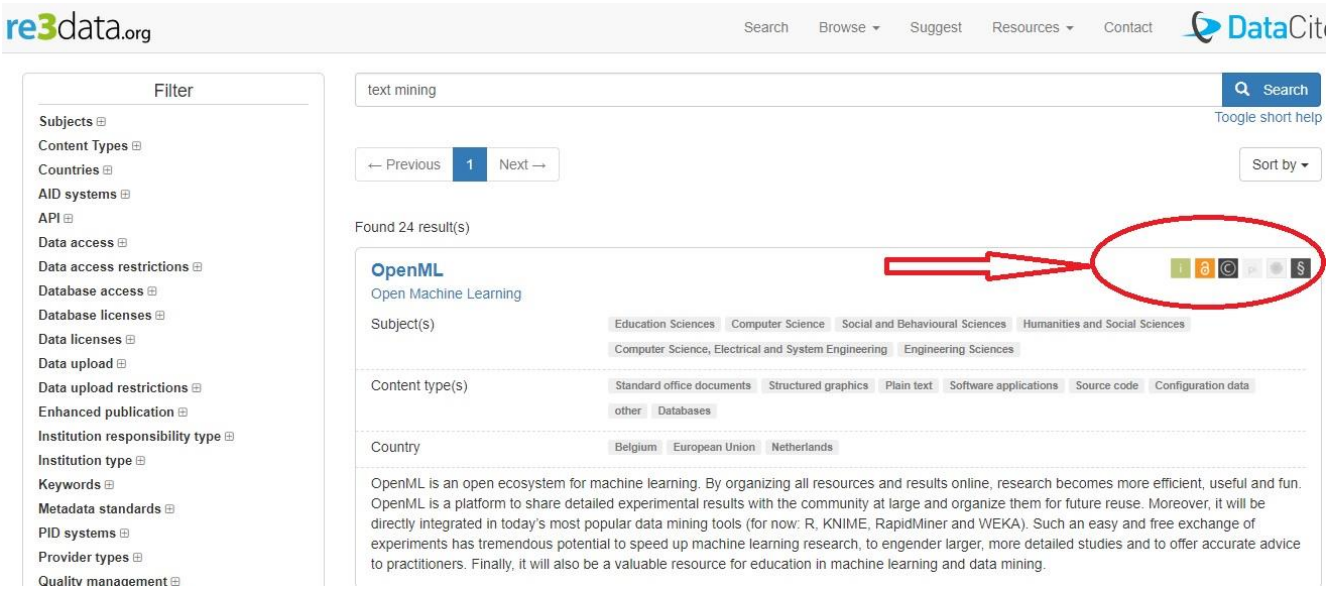


Рисунок 2.22 – Відображення результатів пошуку ресурсу Registry of Research Data Repositories із зазначенням характеристик репозитаріїв.

2.2 Узагальнення результатів аналізу ресурсів наукової та наукометричної інформації

В результаті опрацювання та аналізу систем за запропонованими критеріями було складено узагальнену таблицю. Для узагальнення аналізу було запропоновано тринадцять критеріїв, тобто множина критеріїв складається із 13 компонентів – S = {s1, s2, s3, s4, s5, s6, s7, s8, s9, s10, s11, s12, s3} [87].

Відповідність критеріїв наступна:

- Відкритість системи для користувача – система дотримується політики відкритого доступу або система є передплаченою.
- Необхідність реєстрації на ресурсі для отримання інформації.
- Зазначення інформації про предметну область
- Зазначення інформації про назву публікації/експерименту/проекту
- Зазначення інформації про автора/авторів
- Присутність у результатах пошуку ключових слів
- Наявність реферату/анотації статті/проекту/експерименту
- Можливість переглянути повний текст публікації за результатами пошуку

- Наявність в результатах пошуку посилання на повний текст публікації/опис наукових результатів/експерименту
- Наявність наукометричних показників в результатах пошуку
- Можливість здійснити пошук запроваджено на головній сторінці ресурсу
- Можливість збереження отриманої інформації – результатів пошуку на комп’ютері користувача.

Множина систем налічує 18, відповідність наступна:

- Web of Science
- Scopus
- Google Scholar
- Research Publication Repository
- Joint Research Centre Publications Repository
- Research and Publications Archive
- ArXiv
- Zenodo
- Springer
- Elsevier
- International Organization for Scientific Research
- Scientific Research
- International Journal of Computer
- Open Science Journal
- OpenAire
- Software Heritage
- CERN Scientific Information Service
- Registry Of Research Data Repositories

Отже таблицю відповідності відображено у наступному вигляді (Додаток А).

Якщо припустити, що параметри, шо набувають значення «1» представляють собою безпосередній зв’язок критерія із системою, то чим більше зв’язків – тим кориснішою система буде для користувача і пошук запитуваної інформації

виявиться результативнішим, отже кінцевої мети роботи користувача із системою буде досягнуто.

Висновки до розділу 2

Відповідно до проведеного аналізу, запропонованих критеріїв та систем для аналізу, найбільшу відповідність запропонованим критеріям, що передбачає представлення користувачеві більше можливостей для пошуку, отримання та збереження інформації можна виокремити:

- Наукометричні ресурси, перевагою яких є можливість збереження результатів поуку в окремому файлі для подальшої роботи.
- Архіви та репозитарії, що представляють повноцінні набори інформації про публікацію/проект/експеримент.
- Наукові видання та видавництва наукової літератури.
- Ресурси наукових видань, що представляють докладну інформацію про публікацію, наукометричні показники щодо самого видання, можливість отримання повного тексту документу (згідно політик відкритості окремного ресурсу).

Запропонований підхід до визначення критеріїв для оцінки систем наукової інформації є спробою оцінити повноту представлення результатів пошуку за заданим концептом. Даний підхід може бути розвинений із обґрунтуванням та розширенням кола критеріїв не тільки з точки зору представлення інформації, а й побудови безпосередньо ресурсу.

РОЗДІЛ 3
ОСНОВНІ МЕТОДИ ТА ТЕХНОЛОГІЇ РЕАЛІЗАЦІЇ ЗАДАЧ РОЗШИРЕННЯ МОЖЛИВОСТЕЙ ІНФОРМАЦІЙНИХ СИСТЕМ ДЛЯ НАУКОМЕТРИЧНОГО АНАЛІЗУ

На сьогоднішній день для аналізу текстових масивів, що представлені на веб-ресурсах мережі Інтернет застосовується низка методів, серед яких як методи роботи із текстом так і методи математичної лінгвістики і, зокрема, методи складних мереж, що дозволяють провести відповідні розрахунки та візуалізувати результати.

Для реалізації дослідження джерела текстових даних було обрано англійською мовою і відповідні моделі було розроблено з огляду на роботу із англомовними тексами.

Розглянемо основні методи, що було застосовано для проведення дослідження.

3.1 Методи комп'ютерної лінгвістики.

Із винайденням персонального комп'ютера в середині 20 сторіччя розвиток кібернетики сприяв появі нових наукових напрямків, що виникали на межі наук, що на перший погляд навіть не могли бути пов'язані одна з одною. Таким чином на межі обчислювальної техніки та лінгвістики виникла математична лінгвістика, що на сьогодні ще має назву комп'ютерної лінгвістики.

Основними понятійними категоріями комп'ютерної лінгвістики можна вважати [88]:

- Фрейми – понятійні структури для представлення знань;
- Сценарії – концептуальні структури для представлення знань про стереотипну ситуацію;
- Плани – структури, що надають уявлення про майбутні дії.

Кожен текст містить у собі кілька формальних структур:

- Поверхнева синтаксична структура, де кожне речення може розглядатися окремо;
- Глибинна синтаксична структура – структура, що представляє адекватне відображення ситуації навколишнього світу;
- Семантична структура – виокремлює семантичне значення певних синтаксичних структур.

Комп'ютерна лінгвістика розглядає текст – як письмову форму мовлення, а мову, в свою чергу, як звукову форму тексту.

При комп'ютерному аналізі тексту природньою мовою розглядають наступні рівні:

- Фонетичний рівень;
- Графематичний рівень;
- Морфологічний рівень;
- Синтаксичний рівень;
- Семантичний рівень.

Фонетичний рівень передбачає розгляд звуків з точки зору їх утворення, функцій в мові, фізичних характеристик.

Графематичний аналіз призначений для виокремлення елементів структури тексту: абзаців, заголовків, речень, окремих слів та словосполучень, іноземних лексем, назв, електронних адрес.

Морфологічний аналіз передбачає обробку окремих слів – зокрема виділення основ і флексій, що в подальшому будуть визначати зв'язки між словами.

Синтаксичний аналіз спрямований на розбір більш укрупнених ніж слово – словосполучення та речення. На даному етапі визначається кількість слів у реченні, зв'язок між словами в речення.

Семантичний аналіз спрямований на вирішення задач, що є пов'язані з можливістю визначення значення слова в залежності від контексту, конкретної ситуації, розуміння сенсу фрази чи речення загалом.

Основними задачами комп’ютерної лінгвістики за рахунок проведення вищезазначеного аналізу є:

- Розпізнавання мови та синтез мови, що реалізована на сьогодні у виконанні команд, що представляє користувач голосом, виокремлення окремних слів в мовному потоці, а також інтерпретації мовних повідомлень у вигляді тексту (speech to text). Один із яскравиих прикладів застосування відповідної технології є Google Assistance, що сприймає мову і відповідно формує відповідь на поставлене запитання і підтримує мовний діалог із користувачем.
- Підтримка введення тексу, що на сьогодні є зручним і розповсюдженим інструментом, що використовується в програмах для спілкування - передбачає корегування та підказки щодо правильного введення необхідних лексичних одиниць, а також перевірка введеного тексту.
- Машинний переклад. Якщо перші програми-перекладачі здійснювали переклад з мови на мову за рахунок перекладу кожного слова в реченні, то на сьогодні – це досконалі додатки та он-лайн додатки, що враховують і структуру речення і зв’язок між словами, усталені вирази, тощо. В той же час на сьогодні повноцінно вирішення задачі комп’ютерного перекладу не реалізовано.
- Інформаційний пошук - індексація тексту – виокремлення окремих словоформ, що характеризують заданий текст.
- Компресія тексту є задачею, що сприяє автоматизації реферування тексту – скорочення його викладення та отримання короткого, анотованого.
- Класифікація тексту – задача, що передбачає віднесення кожного документу до відповідного класу, а кластеризація – розподілення масиву документів за тематично близькими рубриками. Задача, що наразі актуальна для формування он-лайн бібліотек, архівів, репозитаріїв.

- Виокремлення фактів та знань зазвичай використовується для обробки потоків інформації стрічок новин інформаційних агенцій для формування аналітичних звітів, тощо.
- Аналіз нормативних текстів передбачає аналіз текстів нормативних документів щодо наявності суперечливих тверджень, логічних пропусків.
- Системи «Питання – Відповідь» вирішують задачу пошуку в текстових корпусах відповіді на поставлене питання.

Одним із основних інструментів комп'ютерної лінгвістики можна вважати словники.

Два основних типи словників, які розрізняють з точки зору призначення – енциклопедичні та лінгвістичні.

Лінгвістичні словники, в свою чергу, поділяються на багато- та двомовні та одномовні. У випадку багатомовних словників мова їде про представленням одних слів співставленням із словами інших мов, одномовні словники – передбачають пояснення слова тією ж мовою.

За функціями словники розрізняють дескриптивні та нормативні. Дескриптивні зображують всі релевантні випадки вживання лексики, нормативні – надають норму вживання слова і надають приклади невірного вживання тих чи інших слів.

Серед одномовних словників виділяють наступні:

- Толкові – містять опис вживання слів;
- Тезауруси – містять поняття однієї галузі знань та сфери діяльності;
- Ідеогрфічні, впорядковані не звично за алфавітом, а за змістом;

Необхідно звернути увагу на такий різновид словників, як частотний [88], в якому лексичні одиниці характеризуються з точки зору частоти уживаності або в наборах текстів, або у мові загалом.

Важливе поняття комп'ютерної лінгвістики – корпус тексту, що означає деякий набір текстів, що пов'язані між собою в логічний спосіб. Важливою властивістю корпусу тексту є репрезентативність – представляє різнобарв'я явища,

що вивчається і представлене корпусом, а також місце цього явища в мовній практиці.

За ступенем організації та структурованості корпуси:

- Електронний архів – тексти на електронному носієві, форма представлення яких не є стандартизована;
- Електронна бібліотека – набір текстів, що представлені відповідно до певного стандарту;
- Корпус текстів – форма текстів є стандартизована і уніфікована.

Параметрами, що застосовувались для проведення аналізу і формування надбудов є інверсна частота терміну та частота терміну.

Деякі слова (поняття) можуть зустрічатися майже в усіх документах певної колекції, і, відповідно не здійснюють впливу на приналежність документу до тієї чи іншої категорії і не може бути ілюстративним для певного документу.

Саме для зниження значущості слів, що зустрічаються майже в усіх документах запропоновано інверсну частоту терміну IDF (inverse document frequency) – це є логарифм відношення числа всіх документів D до числа документів, що містять задане слово – d.

$$IDF = lg\frac{D}{d}, \quad (3.1)$$

Значення параметру менше, якщо слово занадто часто зустрічається в колекції документів.

Параметр частоти терміну – це є співвідношення кількості разів k_i, скільки деяке слово зустрічається в документі, до загальної кількості слів в документі n.

$$TF = \frac{\mathrm{ki}}{\mathrm{n}} \quad (3.2)$$

В рамках проведеного дисертаційного дослідження було опрацьовано інформацію, зокрема текстову з ресурсів Google Scholar, Вікіпедія, Arxiv [89], [90], [91], [92].

При роботі з системами було проведено аналіз текстових даних, що містились на сторінках ресурсів.

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

- для системи Google Scholar - лексичним одиницям, що зображують наукові інтереси – наукові напрямки користувача-науковця, що зазначає їх самостійно у системи при реєстрації та оновленні профілю. Також було сформовано тезауруси, що містили інформації 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 структуровано 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 для запиту концепту.

3.2 Методи статистичного аналізу даних.

Статистичні методи поділяються на два типи:

- одномірні статистичні методи – метод, що застосовується у випадку, коли є єдиний параметр для оцінки елементів, або кожен елемент оцінюється окремо за відповідними параметрами. Одномірні методи класифікуються відповідно до даних, які підлягають аналізу – метричні чи неметричні дані. Неметричні – за номінальною шкалою, метричні – за відносною;

- багатовимірні статистичні методи спрямовані на аналіз даних, для оцінки одного з елементів яких передбачається два чи більше параметрів, як правило використовуються для визначення наявних зв’язків між кількома елементами. Зазвичай, багатовимірний аналіз здійснюється засобами програмних продуктів для проведення відповідних досліджень, але стандартні статистичні інструменти є складовою таких програм, як Excel, Lotus 1-2-3, а також Mathcad.

Основними завданнями статистичного дослідження можна назвати: узагальнення даних та виявлення закономірностей в чітко окреслених умовах часу [93]. Ці дані проявляють себе через подолання випадковості, що притаманна одничним елементам.

В роботі було застосовано багатовимірні статистичні методи, що передбачали оцінку групи елементів – ресурсів наукової інформації за кількома запропонованими параметрами. За рахунок проведення налізу 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.

3.3 Кількісні методи наукометричного аналізу

На сьогодні наукометрія вважається частиною наукознавства і вивчає наукову інформацію із точки зору статистики та динамки її змін у часі.

Серед методів наукометричного аналізу можна зазначити наступні:

- статистичний метод [93] – метод, що використовує кількісні мірки, за його допомогою обраховується кількість журналів, цитувань, тощо;

- метод цитат-аналізу [94] – метод що базується на тому, що цитування джерел є обов'язковим для наукової літератури, базується на такому наукометричному індикаторі, як кількість цитувань. З допомогою цього методу вимірюється кількісний показник наукового доробку вченого – кількість публікацій на посилань на них;
- індекс Гірша, або h-індекс – індекс запропонований фізиком Хорхе Гіршем передбачає комплексну оцінку публікаційного доробку науковця із врахуванням кількості публікацій і посилань на них.
- імпакт-фактор – критерій, запропонований Юджином Гарфілдом, враховує підрахунок цитувань за два роки;
- метод контент-аналізу, що передбачає розподілення тексту на окремі одиниці (наприклад, слова), що потім аналізуються;
- метод тезаурусу – передбачає змістовний аналіз, що проводиться для коректного представлення результатів пошуку серед публікацій.

3.3.1 Індекс Гірша

Індекс Гірша – показник за допомогою якого можна представити комплексну оцінку я кількості публікацій, так і їх цитуванням – кількості і якості наукового доробку вченого.

Індекс Гірша масиву публікацій вченого дорівнює деякому числу h, якщо є h статей із заданого масиву публікацій, кожна з яких була процитована не менше h цитувань, а кожна наступна – не більше h цитувань.

Основна властивість індексу Гірша полягає в тому, що збільшення публікаційної активності без достатньої цитованості робіт не призведе до його збільшення. І навпаки – велика кількість цитувань однієї чи двох робіт серйозно не підвищить значення індексу.

Індекс Гірша було започатковано для авторів, але на сьогодні індекс Гірша обраховується і для колективів вчених, організацій, країн, а також на певні набор публікацій.

Варто зазначити і недоліки індексу Гірша:

- Індекс не враховує предметну область, для вчених, що працюють в різних наукових напрямках індекс буде відрізнятися – це показник, що не є нормалізованим за різними галузями знань;
- Індекс є цілим числом і тому не має розподільної здатності [95];
- Індекс враховує і самоцитування, отже збільшення його значення можливе за рахунок посилання авторами на свої власні публікації.

Популярність індексу Гірша пов'язана в першу чергу із зручністю, адже оцінити науковий доробок організації для адміністраторів науки стало можливим за допомогою одного показника.

3.3.2. Імпакт-фактор

До найбільш розповсюдженого класу індикаторів відносять показники, що оцінюють кількість посилань, що має в середньому одна публікація, що є частиною деякої множини публікацій. Це можуть бути як публікації одного наукового журналу, так і окремого автора чи наукового колективу, організації, країни, тощо.

Зазвичай – це є кількість публікацій, що було оприлюднено за певний проміжок часу, і відповідно необхідно враховувати і термін, протягом якого посилання на ті чи інші статті відповідно до визначеного обсягу публікацій за визначений проміжок часу.

Самим відомим і найбільш розповсюдженим на сьогодні імпакт-індикатором є імпакт-фактор журналу.

При обрахунку імпакт-фактору враховується термін в два роки щодо обсягу публікацій, який розглядається, і рік часу для обрахунку посилань на обраний обсяг публікацій.

Імпакт-фактор характеризує середню кількість посилань, що отримані у звітному році статтями журналу, що були опубліковані попередніх двох роках.

Імпакт-фактор окремого журналу за N рік можна представити у наступному вигляді:

$$\mathrm{IF} = \frac{C(N; N-1,N-2)}{P(N-1,N-2)}, \tag{3.3}$$

де P – загальна кількість публікацій,

P(N) – кількість публікацій за N рік,

C – загальна кількість цитувань,

C(N) – загальна кількість цитувань в N-му році всіх публікацій.

3.4 Теорія графів та теорія складних мереж

Основні проблеми і задачі складних мереж полягають у наступному [96]:

- Дослідження стандартних характеристик графів для складних мереж різної природи – випадкових графів, безмасштабних мереж, мереж малого світу та ін.
- Визначення та дослідження нових характеристик складних мереж – середній та мінімальний шлях, посередництво, коефіцієнт кластеризації.
- Дослідження «фізичних» процесів на складних мережах – дифузії, епідемічних процесів, різних потоків (наприклад, інформації, електричного струму). Також – алгоритм Page Rank – перехід за зв’язками – гіперпосиланнями.
- Методи поновлення, захисту, знищення та оптимізації мереж.
- Пошук наявних зв’язків, що штучно приховані.

Методи, що вирисовуються для вирішення цих задач поділяються на три типи:

- Методи теорії графів – комбінаторні методи;
- Чисельне моделювання;
- Методи теоретичної фізики – від теорії середнього поля до ренорм-групи та діаграмної техніки.

В теорії складних мереж виділяють три спрямування:

- Дослідження статистичних властивостей мереж, що характеризують поведінку мереж;
- Створення моделей мереж;
- Прогнозування поведінки мереж при зміні структурних властивостей.

Складні мережі мають наступні властивості:

- Великі розміри.
- Елементи випадковості при формуванні.
- Ріст та зміни з часом.
- Об’єднання деяких вузлів у групи – ансамблі.

Серед характеристик складних мереж виділяють наступні:

а) параметри вузлів мережі:

- кількість ребер графа, що входять до вузла – вхідна степінь вузла;
- кількість ребер графа, що виходять з вузла – вихідна степінь графа;
- відстань від одного вузла до іншого;
- найбільша з мінімальних відстаней від заданого вузла до інших – ексцентриситет;
- кількість найкоротших шляхів, що проходить через заданий вузол – посередництво;
- загальна кількість зв’язків даного вузла по відношенню до інших – центральність.

б) найкоротший шлях між вузлами визначається як кількість кроків, що необхідно зробити для того щоб за існуючими ребрами дістатися від одного вузла до іншого. При цьому вузли можуть бути з’єднані напряму або опосередковано через інші вузли.

Найкоротшим шляхом є мінімальна відстань між ними. Для всієї мережі може бути застосоване поняття середнього найкоротшого шляху, як середня за всіма парами вузлів мінімальна відстань між ними:

$$l = \frac{2}{n(n+1)} \sum_{i \le j} l_{y,} \quad (3.4)$$

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

в) Коефіцієнт кластеризації – характеризує тенденцію до утворення груп взаємопов'язаних вузлів, так званих клік. Для конкретного вузла коефіцієнт кластеризації показує, скільки найближчих сусідів даного вузла є найближчими сусідами одне для одного.

Відношення реальної кількості зв'язків, що поєднують найближчих сусідів даного вузла i до максимально можливого називається коефіцієнтом кластеризації вузла C_i. Цей показник не може перевищувати одиницю.

Коефіцієнт кластерізації може обраховуватися як для окремого вузла, так і для всієї мережі:

$$C = \frac{1}{n} \sum_{i=1}^{n} C_{j.} \quad (3.5)$$

г) Посередництво – є параметром, що вказує на кількість найкоротших шляхів, що проходять через вузол. Ця характеристика відображає роль даного вузла у встановленні зв'язків в мережі. Вузли із найбільшим посередництвом відіграють головну роль у встановленні зв'язків між іншими вузлами мережі. Посередництво визначається в наступний спосіб:

$$b_m = \sum_{i=j} \frac{B(i,m,j)}{B(i,j)}, \quad (3.6)$$

де B(i, j) – загальна кількість найкоротших шляхів між вузлами і та j, B(i, m, j) – кількість найкоротших шляхів між вузлами і та j, що проходять через вузол m.

За рахунок використання вищезазначених методів стала можливою реалізація запропонованих у роботі підходів до аналізу ресурсів наукової інформації та формування надбудов для виокремлення нових масивів інформації для подальшого аналізу та обробки.

Відповідно до визначених напрямків дослідження дисертаційної роботи було виокремлено ряд задач, реалізація яких передбачає розробку моделей та їх

подальшу реалізацію у вигляді алгоритмів для розширення можливостей існуючих ресурсів наукової інформації для формування нових масивів наукових даних для їх подальшої обробки та аналізу, що виступатиме зручним інструментарієм для науковця при пошуку інформації та розширить коло способів інтерпретації даних, що представляють системи.

3.5 Розширення можливостей отримання нових інформаційних масивів на базі ресурсу Google Scholar.

3.5.1 Побудова мережі предметних областей

Однією задач, реалізованих в рамках дослідження – є побудова моделі предметних областей на базі даних ресурсу Google Scholar. Задача створення моделей предметних областей є актуальною на сьогодні. Адже коло наукових напрямків, що містить кожна предметна область час від часу розширюється з огляду на розвиток науки і технологій.

Моделлю предметної області [97] будемо вважати сформовану спеціальним чином мережу понять – галузеву онтологію. Для створення онтології необхідним є побудова термінологічної основи онтології та виокремлення семантичних зв’зків [98]. В даному дослідженні розглядається модель предметної області, що будується за рахунок зондування великої мережі, зокрема даних системи Google Scholar Citations [65]. Поняття, що беруться для розгляду і побудови необхідної моделі – є теги наукометричного сервісу, що зазначаються кожним автором-науковцем, що реєструється в системі.

Під зондуванням інформаційних мереж будемо розміти виокремлення невеликого обсягу змісту, що представляє інтерес з великих інформаційних мереж, які з технологічних причин не підлягають повному скануванню [96].

На рисунку 3.1 показано інтерфейс сторінки сервісу Google Scholar Citations, на якому зображено результати пошуку за заданим тегом ‘text mining’.

За результатами пошуку відображено перелік зареєстрованих користувачів системи, що позначили заданий для пошуку тег ‘text mining’, як такий, що відповідає наковому напрямку користувача. Наприклад користувач Sophia

Ananiadou зазначила поряд із заданим тегом наступні - Text Mining, Natural Language Processing. Результати пошуку відповідно до цитованості авторів, починаючи з тих, чиї роботи цитуються найбільше – до тих, що найменше [99].

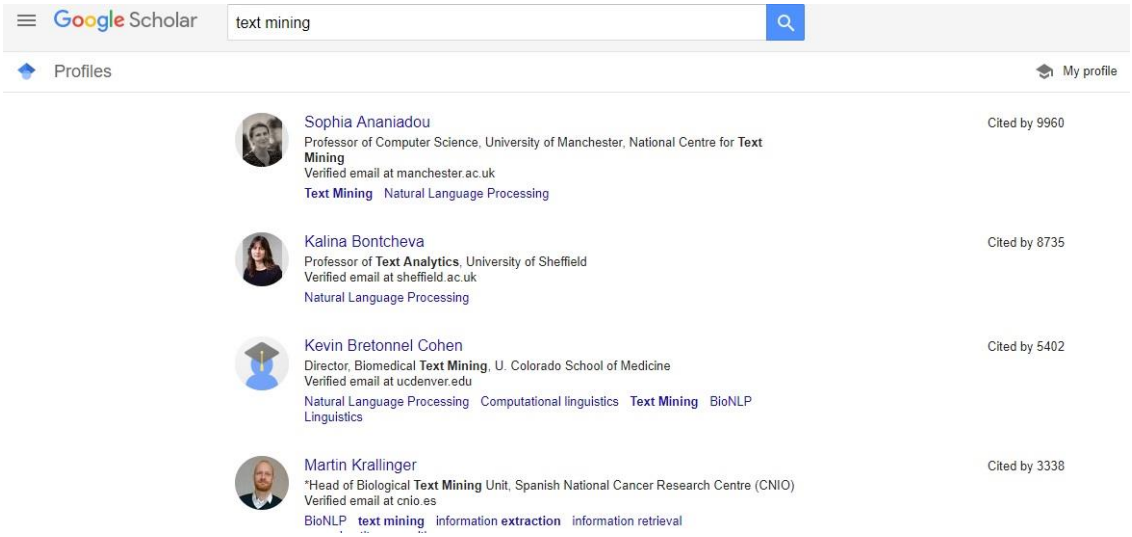


Рисунок 3.1 – Інтерфейс сторінки результатів пошуку Google Scholar Citations

В роботі постає задача опису теоретичних засад і методології автоматизованого формування моделі предметної області, що може бути застосована для будь-якого заданого тегу, що задається для пошуку у системі.

Взаємозв’язок користувачів системи – науковців, авторів наукових публікацій, із відповідними предметними областями можна представити у схематичному вигляді, що зображений на рисунку 3.2.

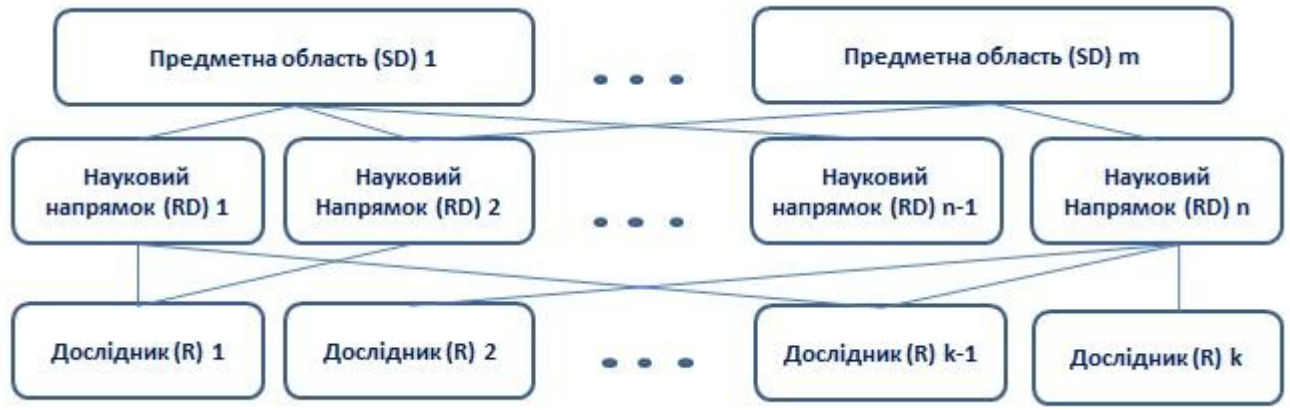


Рисунок 3.2 – Взаємозв’язок користувачів системи Google Scholar Citations із предметними областями

Предметна область може бути представлена у вигляді множини:

$$SD = \{SD1, SD2,\ldots, SDm\},$$

де SDm – конкретна предметна область, якій відповідає певний науковий напрямок – тег, що зазначає користувач системи при реєстрації чи оновленні профілю. Предметна область, та приналежність до неї того чи іншого наукового напрямку визначається експертним шляхом.

Множина, що представляє масив наукових напрямків може бути представлена в наступний спосіб:

$$RD = \{RD1, RD2,\ldots, RDn\},$$

де RDn – науковий напрямок, визначений користувачем. Необхідно зауважити, що один науковий напрямок може бути частиною різних предметних областей одночасно [95], тобто:

$$RDn \in SDn, RDn \in SDn\text{-}2.$$

З огляду на те, що науковий напрямок, зазначений в системі визначає самостійно кожен дослідник, тож зв'язок дослідників із науковими напрямками можна також представити у наступному вигляді:

$$R = \{R1, R2,\ldots, Rk\},$$

де Rk – окремий дослідник, що створив профіль в системі, зазначивши наукові напрямки, що характеризують його наукові інтереси та публікаційні доробки, що представлені в системі, і також дослідник вносить зміни до свого профілю, розширяючи спектр наукових напрямків і відповідно предметних областей, в яких він працює, в рамках розвитку власної наукової кар'єри. Тобто науковець – користувач профілю може різні предметні області, як фахівець з різних галузей знань, так і за рахунок приналежності одних і тих же наукових напрямків різним предметним областям (рисунок 3.3).

Для реалізації поставленої задачі розроблено спеціальний алгоритм, що дозволяє сканувати ресурси сервісу Google Scholar Citations, і в такий спосіб дозволяє отримати репрезентативний набір тегів (позначень понять) як основи майбутньої онтології.

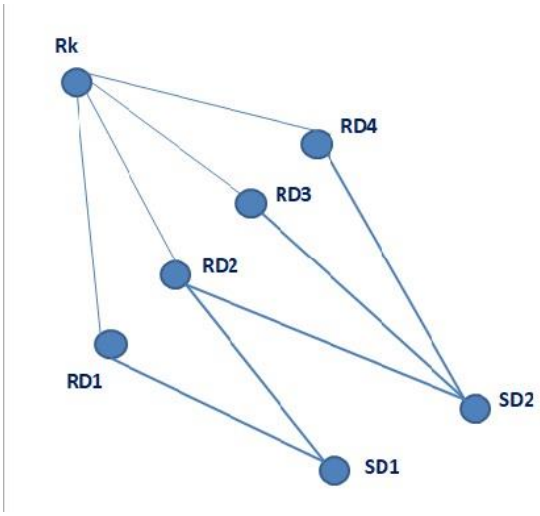


Рисунок 3.3 – Схематичне представлення приналежності одного дослідника кільком предметних областям за рахунок наукових напрямків.

При побудові мереж, що представляють наукові напрямки доцільно застосовувати моделі, що вже були випробувані на піринговых мережах (Peer-to-peer, P2P – рівний з рівним) [96], які засновані на тому, що права учасників є рівні.

У таких мережах відсутні кожен вузол (peer) є як клієнтом, так і сервером. У багатьох випадках P2P є накладеними мережами, що використовують існуючі транспортні протоколи мережі Інтернет. Пірингові мережі містять вузли, кожен з яких взаємодіє лише з певною підмножиною інших вузлів заданої мережі, що відбувається через обмеженість ресурсів.

Якщо розглядати модель інформаційної мережі, яку будемо вважати піринговою мережею, до того ж сполученою з глобальною мережею Інтернет, яка розглядається як зовнішнє середовище. Для реалізації пошуку необхідних даних, у випадку проведено дослідження – це теги, які зазначає користувач системи – науковець, то для таких мереж застосовується кілька моделей. Наприклад, модель, що відповідає методу “широкого первинного пошуку” для мережі, що має розмірності N. Нехай на вході є запит, який з вузла q адресується до всіх сусідів (найближчих за деякими критеріями вузлів). Коли вузол p отримує запит, виконується пошук в його локальному індексі. Якщо деякий вузол r приймає запит (Query) і обробляє його, то він генерує повідомлення-відгук (QueryHit), щоб повернути результат. Повідомлення-відгук включає інформацію про релевантні теги, яка доставляється по мережі вузлу, що запрошує. Інший, “інтелектуальний

пошуковий механізм” (Intelligent Search Mechanism, ISM) забезпечує поліпшення швидкості і ефективності пошуку інформації за рахунок мінімізації витрат на кількість повідомлень, що передаються між вузлами, та мінімізації кількості вузлів, які опитуються для кожного пошукового запиту [99]. При цьому для досягнення поставленої мети – проводиться оцінка лише тих вузлів, що найбільш відповідають запиту. В даній роботі буде зображено саме модель, що є близькою до ISM.

Процес зондування опорної мережі здійснюється за алгоритмом, зображеним на рисунок 3.4:

Рисунок 3.4 – Алгоритм зондування опорної мережі

Згідно алгоритму відбувається реалізація наступних кроків:

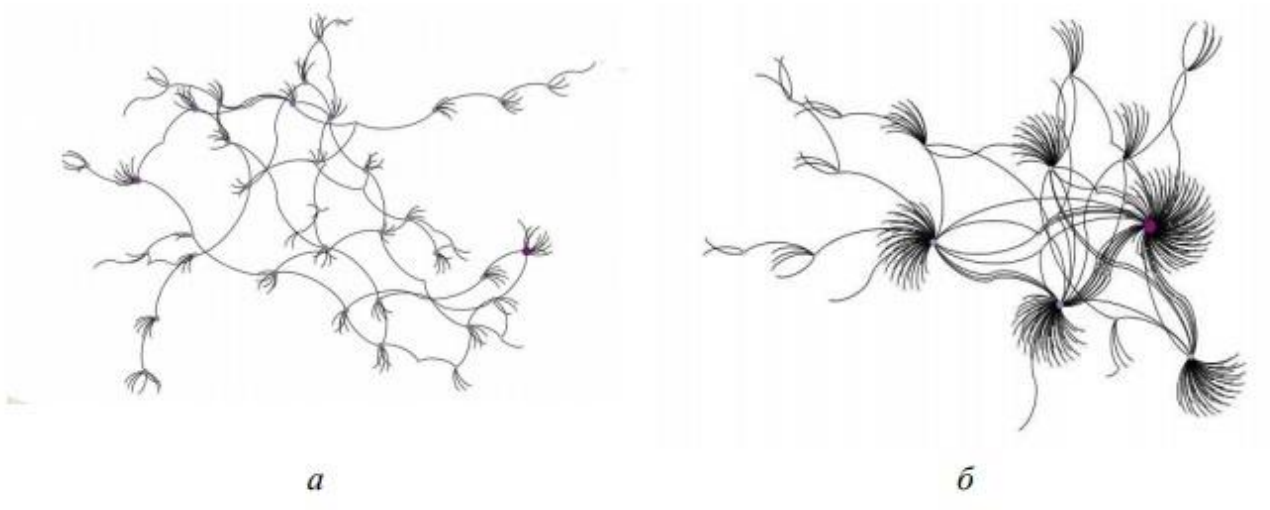
Зондування опорної модельної мережі здійснюється за таким алгоритмом:

- Спочатку обирається визначена кількість вузлів мережі, що зондується. Вузли визначаються як базові для нової мережі, що відповідає результатам зондування.
- Для кожного з базових вузлів визначаються сусідні вузли, які додаються до мережі, що створюється.
- Здійснюється перехід до сусіднього вузла мережі, що має найбільшу ступень.

- Якщо вибирається вузол, до якого вже було здійснено перехід за цим алгоритмом, здійснюється перехід до наступного базового вузла з початкового переліку і здійснюється перехід до пункту 2.

5. Якщо перелік базових вузлів завершено, будемо вважати, що мережу, що відповідає результатам зондування, побудовано.

Наведений алгоритм було застосовано для двох найпоширеніших Ердеша-Рені(ER) і Барабаші-Альберта (BA) (рисунок 3.5).



а *б*

Рисунок 3.5 – Приклади мереж, побудованих зондуванням модельних мереж, де: а – Ердеша-Рені; б – Барабаші-Альберта.

Модель Ердеша-Рені – є випадковою мережею, що будується в наступний спосіб: N спочатку не з’єднаних вузлів попарно поєднуються з певною ймовірністю p. В результаті створюється мережа з pN N(1) 2 − випадково вибраними зв’язками.

Модель Барабаші-Альберта – одна з моделей мереж із степеневим розподілом ступенів вузлів – безмасштабних мереж. Ця модель враховує і зростання мережі (динаміку), і принцип переважного приєднання, який полягає в тому, що чим більше зв'язків має вузол, тим переважніше для нього створення нових зв’язків, що знову утворюються вузлами. Вузли з більшим ступенем мають більшу вірогідність приєднання нових вузлів – відповідно створення нових зв’язків. Необхідно зазначити, що безмасштабні мережі на сьогодні є найбільш популярними, такі як

веб-простір із гіперпосиланнями, соціальні мережі, мережі слів у літературних творах, тощо. Мережі понять, які природно формуються учасниками мережевих сервісів ймовірно також мають властивість безмасштабності, але не завжди це можна перевірити при наявності повної інформації. Якщо мережа є складною і великою, як Google Scholar Citations, може бути використаний саме алгоритм зондування. Необхідно зазначити, що результати будь-якого зондування не завжди будуть адекватно відображати природу великої мережі, що досліджується. Є фактори, що залежать від алгоритму цієї процедури. Візуально якісні результати зондування мереж Ердеша-Рені та Барабаші-Альберта з близькими параметрами (1000 вузлів, понад 2000 зв'язків) наведені на рисунку 3.5. За рахунок порівняння можна зауважити, що зв'язані області (гілки), що відповідають окремим поняттям, у першому випадку досить довгі, а вузлів, за якими йде маршрут зондування, більше, ніж у другому, більш цікавому, випадку. В дослідженні не є важливими чисельні результати і параметри мереж, важливою є оцінка вигляду зв'язаних ланцюжків, що моделюють гілки понять. Необхідно зуважити, що реальним мережам притаманний ще й феномен "клуба багатіїв" (Rich Clube), який обумовлює щільнішу зв'язність найбільших вузлів. Тому наведений алгоритм має таку особливість, як швидке зациклювання, що приведе до ще більшого скорочення гілок понять.

Саме за рахунок якісного моделювання було доведено що можливість формування невеликих зв'язаних гілок, які відповідають поняттям, що цікавлять користувачів сервісу Google Scholar Citations.

Імплементацію наведеного алгоритму було адаптовано до реальної мережі понять в спосіб, що більш розширений порівняно із алгоритмом, наведеним вище:

- експертним шляхом визначається перелік базових тегів, які визначаються авторами при створенні та редагування профілю і є заданими поняттями для побудови мережі.
- обирається тег з визначеного експертами переліку.

- відкриваються сторінки веб-сервісу, що відповідають цьому тегу – максимальна кількість таких сторінок обмежується визначеним параметром.
- до мережі додаються усі теги, що містяться на обраних сторінках.
- здійснюється перехід до сторінок, що відповідають тегу, що найбільше повторювався на сторінках.
- якщо вибирається тег, до якого вже було здійснено перехід за цим алгоритмом або “відхід від теми” (виявляється за результатом змістовного аналізу).
- якщо перелік базових тегів завершено, мережа вважається побудованою. Інакше здійснюється перехід до перехід до пункту 2 – наступного базового тегу з початкового переліку.

У запропонованій моделі предметної області застосовуються зв’язки між областями професійних інтересів окремих вчених. Розглядається компактифікація біографа “вчений – галузі науки, що його цікавлять”. Ці зв’язки дозволили припустити наявність загального наукового апарата та семантичний зв’язок. Також запропоновано та реалізовано підхід формування моделі предметної області, при формуванні якої застосовуються знання, заздалегідь зазначені вченими, що є користувачами системи Google Scholar Citations. Запропоновану модель можна застосовувати для різних галузей науки.

Запропонований підхід може бути розширено за рахунок врахування інших даних. Що містить система: інформація, про публікації, яка передбачає зазначення назви наукового видання, що в свою чергу дозволить сформувати масив наукових журналів відповідно до наукових напрямків.

3.5.2 Побудова мережі співавторів на базу ресурсу Google Scholar Citations

Вивченню мереж співавторів, так само як і сервісу Google Scholar Citation (http://scholar.google.com/citations), присвячено велику кількість досліджень, що, в свою чергу, підтверджує актуальність проведених досліджень. Серед них методи

побудови мереж співавторів, визначення значущих вузлів, структури мережі, дослідження цитування в Google Scholar, а так само відповідних корпусів [97].

Як і для побудови мереж предметних областей, для побудови мережі співавторів використовуються теги, що задані користувачами сервісу Google Scholar Citations. Для побудови мережі співавторів було запропоновано відповідний алгоритм [100], [101] що може бути адаптований до різних предметних областей.

Алгоритм побудови мережі співавторів представлений на рисунку 3.6.

Рисунок 3.6 – Алгоритм побудови мережі предметних областей на базі ресурсу Google Scholar Citations

Алгоритм передбачає наступні кроки:

- обирається перший автор (вузол), з якого починається зондування.
- експертним шляхом визначається перелік базових тегів-дескрипторів, що відповідають поняттям заданим для пошуку.
- відкривається сторінка, що відповідає обраному автору.
- до створюваної мережі додаються всі співавтори, що містяться на сторінці обраного автора. Формуються ребра-зв’язки до цих вузлів (співавторів) з вихідного вузла (автора).

- з переліку вузлів мережі, що формується, випадковим чином обирається автор, на сторінку якого планується перехвд для подальшого аналізу. Тематика автора має відповідати тематиці обраної предметної області (його теги входять до складу тегів-дескрипторів, визначених на кроці 2, і не входять до складу тих вузлів, до сторінок яких вже був здійснений перехід.
- якщо такий вузол-автор обрано, то здійснюється перехід до кроку 3, якщо такого вузла не існує – мережа вважається побудованою, а зондування закінченим.

Відповідно до запропонованого алгоритму процес зондування мережі, починаючи з певного вузла, припиняється при “зациклюванні”, тобто коли відповідно до алгоритму має відбуватися перехід до вже пройденого вузла, а також при відхиленні сусідніх вузлів від основної тематики.

Застосування методів кластерного аналізу дозволяє виявляти найбільш тісно пов’язані між собою групи вчених-співавторів.

Необхідно зазначити, що існуючі моделі автоматичного формування мереж співавторів базуються на особистій участі експертів при виборі вузлів і зв’язків.

В запропонованій моделі для побудови мережі співавторів використовуються лише базові теги. Надалі алгоритм передбачає використання знань, що закладені співавторами, теги позначені як головні для них – експертне середовище в цьому випадку істотно розширюється. Необхідно зазначити, що система Google Scholar Citations є зручною щодо доступу до інформації, не передбачає створення власного профілю користувача для доступу до інформації, доступ є необмежений. В той же час кожен зареєстрований має можливості збереження отриманої і скорегованої інформації. Пошук в системі Google Scholar є спрощеним, на відміну, наприклад, від сервісу Science Direct [58], що пропонує користувачеві одразу кілька параметрів для здійснення пошуку. Можливості уточнення пошуку за тегом є більш орієнтованими на користувача і дозволяють за рахунок чітких уточнень швидше отримати необхідну інформацію.

3.6 Побудова нових інформаційних масивів на базі енциклопедичного ресурсу Wikipedia.

3.6.1 Побудова онтології

З огляду на те, що на сьогодні Wikipedia (http://wikipedia.com) – є найбільш популярним он-лайн енциклопедичним ресурсом відкритого доступу – і в той же час щоденним інструментом широкого кола користувачів, що черпають не тільки буденну тлумачну інформацію, а й елементарні і більш складні наукові визначення та отримують уявлення про наукові явища та поняття.

В роботі запропоновано методику побудови онтології на базі автоматичного моніторингу і аналізу мережевих інформаційних ресурсів довідкового характеру. В роботі розглядається мережа понять, що відповідають термінам-заголовкам статей мережевої енциклопедії Wikipedia та гіперпосиланням, що містяться у тексті публікацій.

Закономірно, що мережа понять, яку можна побудувати за рахунок сканування енциклопедичного сервісу матиме великі розміри, але вона може бути обмежена певною тематикою, яка відповідає заданій предметній області. Зазначена властивість, в свою чергу, ускладнює сприйняття побудованої мережі і призводить до ефекту, що має назву – зсув тематики (Topic Drift). Для уникнення зазначеного ефекту застосовується тематична фільтрація – для аналізу використовуються лише ті статті з Вікіпедія, які містять базовий термін, що визначається експертом на початку роботи. Відповідність цим дескрипторам визначає розмір сформованих мереж – моделей предметних областей, а також динаміку їх формування. Також розпізнавання кластерів в таких мережах може розглядатися як основа для виокремлення визначених наукових напрямків [102].

Енциклопедичний ресурс було обрано для розгляду з огляду на його доступність – зокрема доступність формування надбудов. Ресурс є абсолютно відкритий для користувачів, не потребує передплати та додаткової реєстрації для доступу до інформації. Статті, що містить ресурс формуються та редагуються авторами-дописувачами із зазначенням посилань на відповідні джерела, що використовуються для формування текстових масивів із різноманітної тематики.

Для роботи із системою застосовуються базові теги – концепти для пошуку інформації – статті, що відповідає заданому тегу (рисунок 3.7).

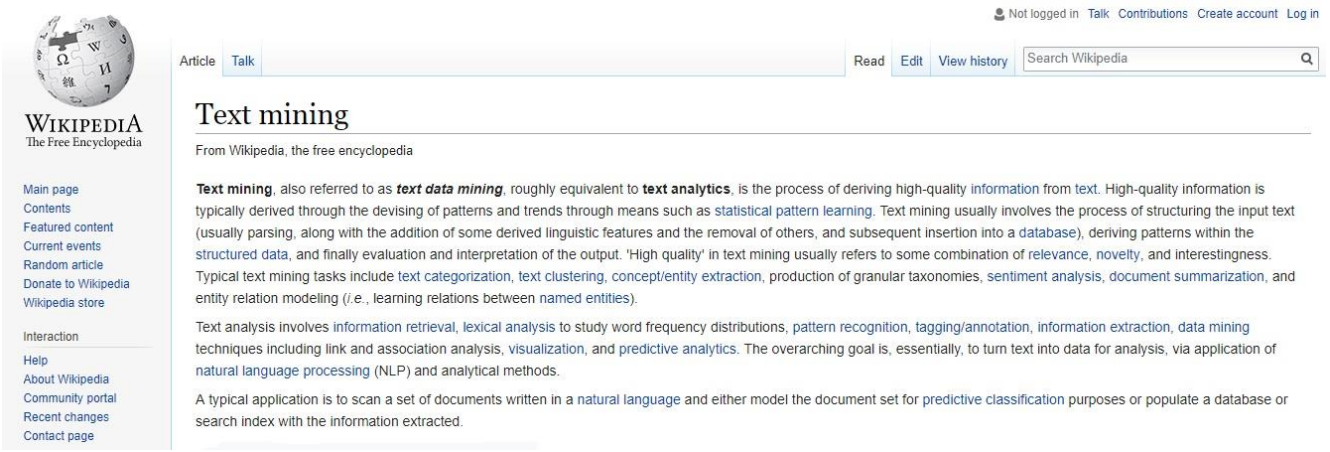


Рисунок 3.7 – Інтерфейс користувача системи – стаття, що відповідає тегу для пошуку ‘text mining’

Зважаючи на базові теги, що відповідають певній предметній області, визначено представлення інформації в цій системі. Також було визначено, що вільний перехід за гіперпосиланнями призводить до ефекту зсуву тематики.

Для реалізації задачі побудови мережі предметних областей за даними сервісу Вікіпедія із врахуванням можливого ефекту зсуву тематики запропоновано алгоритм, що відображено на рис. 3.8.

Відповідно до запропоновано алгоритму передбачається виконання наступних кроків:

- обирається термін (поняття) для пошуку;
- відкривається сторінка веб-сервісу (стаття Вікіпедія), що відповідає обраному терміну.

До створюваної мережі додаються всі терміни-поняття, що відповідають гіперпосиланням на обраній сторінці.

Формуються ребра-зв’язки до цих вузлів з вихідного вузла.

Алгоритм роботи програми зображено на рисунку 3.8 та передбачає наступні дії:

- статті визначаються як базові, якщо вони містять гіперпосилання на статтю, що відповідає першому терміну-поняттю, що було задано для пошуку.

Рисунок 3.8 – Узагальнений алгоритм побудови мережі предметних областей на базі ресурсу Wikipedia.

- із списку вузлів мережі, що формується, визначається той, за яким ще не здійснювалося переходу, на сторінку якого планується перейти для подальшого аналізу. Цей вузол має відповідати вимогам пункту 3 та не має входити до складу тих вузлів, до сторінок яких вже був здійснений перехід.
- якщо такий вузол обрано, то здійснюється перехід до пункту 2.
- якщо такого вузла не існує, то вважається, що мережу, що відповідає моделі предметної області, побудовано.

Згідно алгоритму процес збирання інформації зі статей Wikipedia, починаючи із заданого поняття, припиняється, коли відповідно до алгоритму вже неможливий перехід до нового вузла (базових вузлів для переходу вже не лишається), тобто “зациклювання” неможливе.

У роботі запропоновано і реалізовано алгоритм формування моделей предметних областей за рахунок автоматичного аналізу енциклопедичного ресурсу Вікіпедія. Лістінги програм представлені в додатку Б. Від статичних моделей предметних областей такий підхід відрізняється урахуванням динамічної зміни контенту бази даних цього сервісу, урахуванням нових понять та нової інформації, що додається до існуючих статей за рахунок редагування та доповнення інформації різними авторами - користувачами Вікіпедія.

3.6.2 Індекс популярності автора на базі енциклопедичного ресурсу Wikipedia (Вікі-індекс)

Варто підкреслити, що енциклопедичний ресурс Вікіпедія – є доступним і популярним сервісом мережі Інтернет. Ресурс дозволяє не тільки поновлювати, редагувати та наповнювати його всім користувачам, але й розробляти додаткові інструменти для роботи з інформацією системи.

В роботі запропоновано новий індекс – Wikipedia Index – WI , що дозволить оцінити популярність авторів наукових публікацій і розраховується за рахунок аналізу інтернет-енциклопедії Вікіпедія [103], [104].

На відміну від існуючих індексів цитування, цей індекс Вікіпедії дозволяє оцінити не тільки індекс популярності як при розрахунку загального числа цитувань або специфічний індекс Гірша, що частіше використовується для оцінки наукового рівня авторів.

Запропонований показник дозволяє оцінити популярність автора, його видимість та певний вплив в найбільш популярному в Інтернеті осередку – в енциклопедії Wikipedia.

В роботі запропоновано розраховувати цей індекс тільки в межах предметної області – це з одного боку дозволить уникнути помилкового підрахунку для авторів

із однаковими прізвищами – однофамільців, а з іншого боку – для забезпечення повноти охоплення предметної області.

В роботі запропонований алгоритм і технологія підрахунку Wikipedia Index на базі зондування мережевої енциклопедії.

Сьогодні існує певний спектр показників, що є інструментраєм, який використовується для проведення наукометричних досліджень. Ці показники орієнтовані на те, щоб надавати інформацію щодо рівня наукових здобутків вчених, вплив їх наукового доробку на суспільство та їх впливовість у світовому науковому просторі. Якщо найбільш простим індексом є кількість публікацій автора, то такий показник не відображає якісних параметрів, що відображаються за рахунок кількості цитувань. Але кількість цитувань не є ілюстративним, якщо мова йде про продуктивність автора, тому що автор всього однієї, але дуже популярної роботи може перевершити за цим показником вчених, що регулярно публікують свої результати.

Індекс Гірша, що обраховується кількома наукометричними ресурсами. Принцип його обрахунку достатньо простий, при цьому він містить як кількісну складову – кількість публікацій вченого, так і якісну складову – кількість цитувань цитувань автора. Індекс обраховується на базі розподілення цитувань робіт даного дослідника. Але індекс Гірша орієнтований на наукову важливість, вагомість автора не повно відображає суспільну важливість результатів, що отримані автором. Показник є сугубо науковим, обраховується ресурсами, що містять виключно наукову інформацію – Web Of Science, Scopus, Google Scholar Citations.

В той же час суспільна складова науки є достатньо великою і для такої оцінки необхідно використовувати науково-популярні, соціально-значущі джерела. Як один із підходів до розв’язання зазначеної задачі запропоновано методологію розрахунку нового індексу – Вікі індексу популярності автора.

Однією із найважливіших переваг показника є те, щоу до розгляду береться ресурс Вікіпедія – найбільший і найдемократичніший в мережі Інтернет. Доступ до ресурсу не передбачає передплати, крім того енциклопедія є доступною для скачування у повному обсязі.

Для первинного доступу до цієї системи було застосовано спеціальні терміни – імена вчених і терміни з цільової проблематики, за якими існують відповідні статті, що створюються, редагуються та доповнюються експертами-авторами.

Вивченню моделей предметних областей, так само як і сервісу Вікіпедія (http://wikipedia.com), присвячена велика кількість робіт, що підтверджує актуальність проведених досліджень [105]. Серед них, зокрема, методи побудови мереж співавторів, визначення значущих вузлів, структури мережі, дослідження цитування, а так само відповідних корпусів [106].

Вікі-індекс поплуряності автора обраховується у спосіб зазначений нижче.

Припустимо, що бібліографічні посилання на автора зустрічаються в N статтях ресурсу Вікіпедія.

Відсортовані за зменшенням ряд із кількості згадувань автора в біблографічній частині цих статей позначимо: R_1, R_2, ..., R_N.

Вікі-індекс популярності автора (WI) відповідає максимальній кількості статей (WH) з Wikipedia, в яких кількість бібліогарфічних посилань на цього автора не менше значення WH, що є помноженим на деяку нерегресну та невід'ємну функцію від N, тобто (в даному випадку розглянуто корінь квадратний):

$$WI = WH \times \sqrt{N} = \max\left(i : R_i > i\right) \times \sqrt{N}$$

Вікі-індекс популярності автора (науковця) близький до індексу Гірша, однак, він враховує не кількість публікацій, які посилаються на публікації автора, а кількість бібліографічних посилань на роботи автора і кількість статей з Wikipedia, що містять ці посилання.

Крім цього варто звернути увагу на множення на функцію від N, що відображає врахування популярності та забезпечує великий розкид значень для різних авторів.

Слід відмітити, що індекс популярності автора має бути прив'язаний до його предметної області, з одного боку – для того щоб уникнути помилкового підрахунку для авторів із однаковими прізвищами, а з іншої – для забезпечення повноти охоплення для всієї предметної області.

Вікі-індекс популярності, в свою чергу може стати ілюстративним при розгляді конкретного дослідника і його видимість в рамках ресурсів із широким колом користувачів. Також є зручним інструментом для визначення обсягу інформації про певну організацію та її дослідників у відкритих і доступних користувацькому загалу ресурсах, що в свою чергу розширить коло пошукових результатів для тих, хто є зацікавлений тематикою і конкретними дослідженнями тих чи інших організацій. На сьогодні, коли коротка і доступна інформація – це перший крок до формування зацікавленості організацій та окремих дослідників у пошуку партнерів для формування колаборацій.

3.7 Побудова мережі предметних областей на базі ресурсу препринтів arXiv.

arXiv є одним із популярних ресурсів для розміщення результатів досліджень – найбільший архів електронних публікацій та їх препринтів відкритого доступу.

Репозитарій було запроваджено у 1991 році. Роботу ресурсу було спрямовано на розміщення публікацій підготовлених до друку за напрямком «Фізика», але на сьогодні ресурс постійно розширюється і додаються нові роздлиі та відповідні підрозділи з інших наукових напрямків.

arXiv – є допоміжним інструментарієм для науковців в усьому світі. Ресурс є актуальним інструментом для користувачів з країн із обмеженим доступом до наукової інформації за рахунок можливості користування дзеркалами.

До сьогодні всю увагу до ресурсу та досліджень на базі нього було зосереджено на способах виявлення плагіату [109] та впровадженні протоколів відкритого доступу на базі архіву препринтів [110].

В той же час arXiv представляє собою базу наукових препринтів та опублікованих праць за різними науковими напрямками, що оновлюється щоденно і містить найновіші результати досліджень з різних галузей знань.

Ресурсом передбачено процедура схвалення (endorsement) статті перед опублікуванням із залучення експертів з різних наукових напрямків.

Доступність ресурсу всім користувачам мережі Інтернет дає можливість застосовувати відповідні моделі для розробки та реалізації алгоритмів для отримання нових масивів інформації на базі ресурсу та подальшої інтерпретації отриманих результатів.

Основною задачею дослідження є розробка моделі «Концепт – система наукових напрямів» ресурсу препринтів arXiv для розробки, реалізації алгоритму для отримання інформації щодо публікацій за заданим концептом. За результатами пошуку відтворити у вигляді схематичних зображень та провести оцінку результатів пошуку і подальшу інтерпретацію результатів.

Під концептом розуміємо [111] те, що називає зміст поняття, в даному випадку – зображення і зміст наукового явища, його характеристика, ознака. В той же час заданий для пошуку концепт може представляти собою не тільки слово або словосполучення, що визначають науковий термін, характеристику наукового процесу, тощо, а й власні імена та назви.

Мережа предметних областей – це спосіб представлення моделі предметних областей за рахунок визначення узагальнених описів предметної області, представлених власне їх назвою та назвами підпорядкованих їй структурних одиниць – наукових напрямків, що більш конкретно описують предметну область, які визначені інформаційною системою, на базі якої будується задана мережа, або на базі запропонованої систематизації предметних областей.

На сьогодні автором запропоновано реалізацію задачі побудови мережі предметних областей та дерева понять [112] на базі заданого концепту. Зображені у статті підходи розширюють шляхи інтерпретації отриманих результатів за рахунок апробації на окремому концепті.

Для реалізації поставленої задачі необхідно виокремити перелік наукових напрямів, визначених самою системою і перелік піднапрямків, що деталізують кожен науковий напрям.

Архів передбачає 8 розділів – наукових напрямів, за якими розподілені публікації:

- Computer Science (41 піднапрямок)

- Economics (1 піднапрямок)
- Electrical Engineering and System Science (3 піднапрямки)
- Mathematics (32 піднапрямки)
- Physics (51 піднапрямок)
- Quantitative Biology (10 піднапрямків)
- Quantitative Finance (10 піднапрямків)
- Statistics (6 піднапрямків)

Для кожного наукового напряму було сформовано словник (додаток Б), що містить назву наукового напряму і має вигляд таблиці (таблиця 2) , що складається з трьох полів із наступною інформацією:

Таблиця 3.1 – Частина словника «Computer Science»

N п/п	**Скорочена назва піднапрямку**	**Повна назва піднапрямку**
1.	cs.AI	Artificial Intelligence
2.	cs.AR	Hardware Architecture
3.	cs.CC	Computational Complexity
4.	cs.CE	Computational Engineering, Finance and Science

Така необхідність виникла з огляду на те, що 17 квітня 2018 року ресурсом було запроваджено новий пошуковий інтерфейс, який передбачає скорочене зазначення назви піднапрямку (рисунок 3.9) в переліках результатів пошуку, і на відміну від попереднього представлення результатів виникла необхідність тлумачення скороченої назви піднапрямку для подальшої коректної візуалізації результатів.


5. arXiv:1803.09617 [pdf] cs.IT

Correlation properties of signal at mobile receiver for different propagation environments

Authors: Cezary Ziolkowski, Jan M. Kelner, Leszek Nowosielski

Abstract: An issue of the parameter selection in various branches of a multi-antenna receiver system determines its effectiveness. A significant effect on these parameters are correlation properties of received signals. In this paper, the assessment of the signal correlation properties for different environmental conditions is presented. The obtained results showed that depending on the receiver speed, the... ▽ More

Submitted 26 March, 2018; originally announced March 2018.

Comments: 4 pages, 5 figures

MSC Class: 94A40; 94A05; 94A12; 94A17 ACM Class: E.4; H.4.3

Journal ref: , Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, 20-24.02.2018., pp. 1-4

Рисунок 3.9 – Інтерфейс сторінки результатів пошуку на ресурсі препринтів arXiv

Робота розробленої системи передбачає задання концепту для пошуку, і за результатами пошуку виокремлення наукових напрямків та піднапрямків, в рамках яких було здійснено дослідження і опубліковано результати на ресурсі і подальше представлення результатів.

Також для масиву публікацій було застосовано додатковий пошук, що виокремлює публікації, що були опубліковані в наукових виданнях, і відповідно відібрано і наукові напрямки.

Модель «Концепт – система наукових напрямків» можна представити у вигляді схеми (рисунок 3.10), що представляє масив наукових напрямків і піднапрямків, що формують словник і безпосередньо зв'язок концепту із цими параметрами.

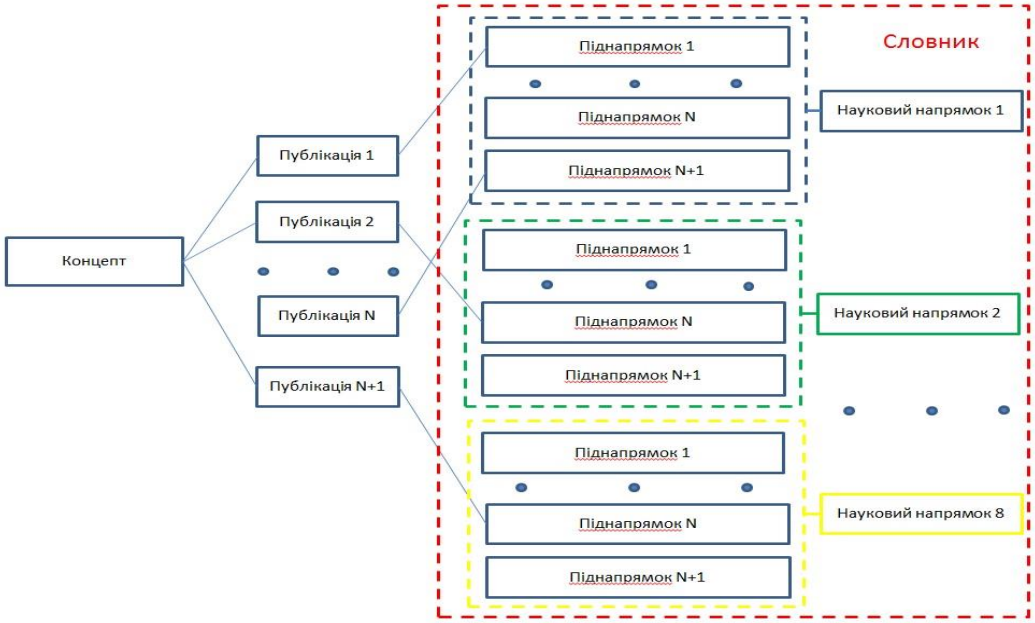


Рисунок 3.10 – Модель «Концепт-система наукових напрямків»

Алгоритм аплікації моделі передбачає виконання кроків зображених на рисунку 3.11.

Для заданого концепту, що може бути представлений одним словом чи словосполученням відбується пошук масиву публікацій, в яких він відображений. Алгоритм побудови мережі предметних областей для заданого концепту передбачає визначення предметних галузей та наукових напрямків, для яких заданий концепт є притаманний. Реалізація алгоритму здійснюється за рахунок обробки інформації, що є результатом пошуку.

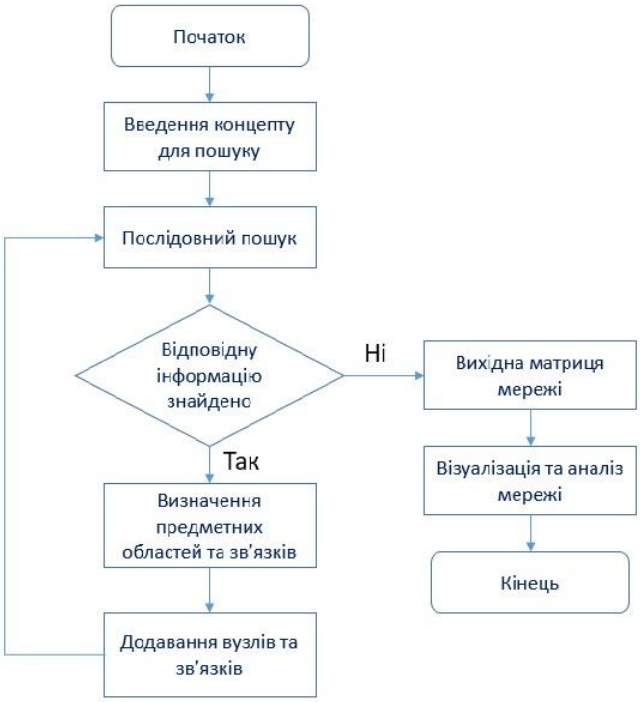


Рисунок 3.11 – Узагальнений алгоритм побудови мережі предметних областей на базі ресурсу препринтів arxiv

Для реалізації поставленої задачі було складено 8 словників, що відповідають науковим напрямкам – предметним областям представленим на ресурсі arXiv. Словники та їх зміст містяться у Додатку В.

Вся робота здійснювалася на базі реферативної інформації - результатів пошуку для заданого концепту.

Алгоритм побудови мережі, що охоплює предметні галузі до яких застосовується задане поняття передбачає виконання наступних кроків:

а) вершиною мережі є вузол, шо тотожний із концептом, що було задано для пошуку.

б) для заданого концепту було отримано перелік реферативної інформації, що містить наступні дані (рис. 3.12) і буде підлягати подальшій обробці:

- Номер по порядку, ідентифікатор в системі, що має наступний вигляд: **arXiv:** XXXX.XXXXX [***], де XXXX.XXXXX – номер публікації в системі, *** - перелік доступних форматів файлів для завантаження;
- Назва публікації;
- Автор (и);

- **Comments** – поле, що, як правило, містить інформацію, про кількість, сторінок публікації, кількість малюнків та інших елементів (для деяких публікацій дане поле може бути відсутнє);
- **Journal-ref** – поле, що містить інформацію про видання, в якому розміщена дана публікація (є в наявності для публікацій, що вже є опубліковані);
- **Subject** – поле, що містить назву предметної області або конкретизованої інформації щодо наукового напрямку в рамках предметної області (відповідно до того, в який спосіб автор публікації зазначив при поданні публікації для розміщення її на ресурсі).

в) для кожної публікації виокремлюється назва наукового напрямку, що зазначена в отриманій реферативній інформації (рисунок 3.12).

г) назва наукового напрямку, зазначена в реферативній інформації для публікації, порівнюється із усіма словниками предметних областей. Назва предметної області, в словнику якої було знайдено назву наукового напрямку – є наступний вузол графу, що з'єднаний із вершиною – заданим концептом.

д) наступний вузол – назва наукового напрямку, що було виокремлено в результаті роботи із реферативною інформацією. Цей вузол з'єднаний з вузлом, що позначає предметну область, яку деталізує отриманий науковий напрямок (рисунок 3.13).

е) відбувається перехід до наступного результату пошуку.

ж) якщо для наукового напрямку вже було побудовано вузол – назву предметної області, то будується тільки вузол – назва наукового напрямку, що з'єднується із вузлом – відповідною предметною областю.

и) якщо для назви наукового напрямку вже було побудовано відповідний вузол, то відбувається перехід до п.6. Якщо для назви наукового напрямку побудову відповідних вузлів ще не було здійснено відбувається перехід до п.4.

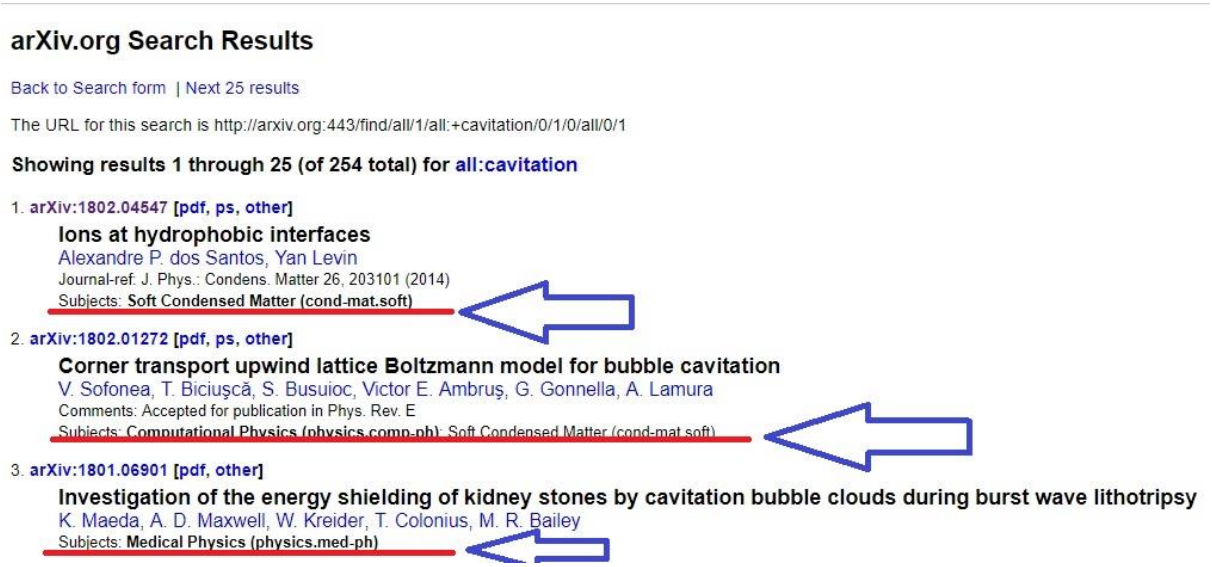


Рисунок 3.12 – Представлення результатів пошуку із позначенням рядка, що містить інформацію про приналежність до наукового напрямку

Якщо назву наукового напрямку, зазначеної в полі “Subject” за результатами пошуку було знайдено в двох словниках, то будується два вузли – назви предметних областей, або той вузол, який ще не було побудовано, і відповідно з'єднаний з ними обома вузол – назва наукового напрямку.

Мережа вважається побудованою по завершенні сканування всіх результатів пошуку.

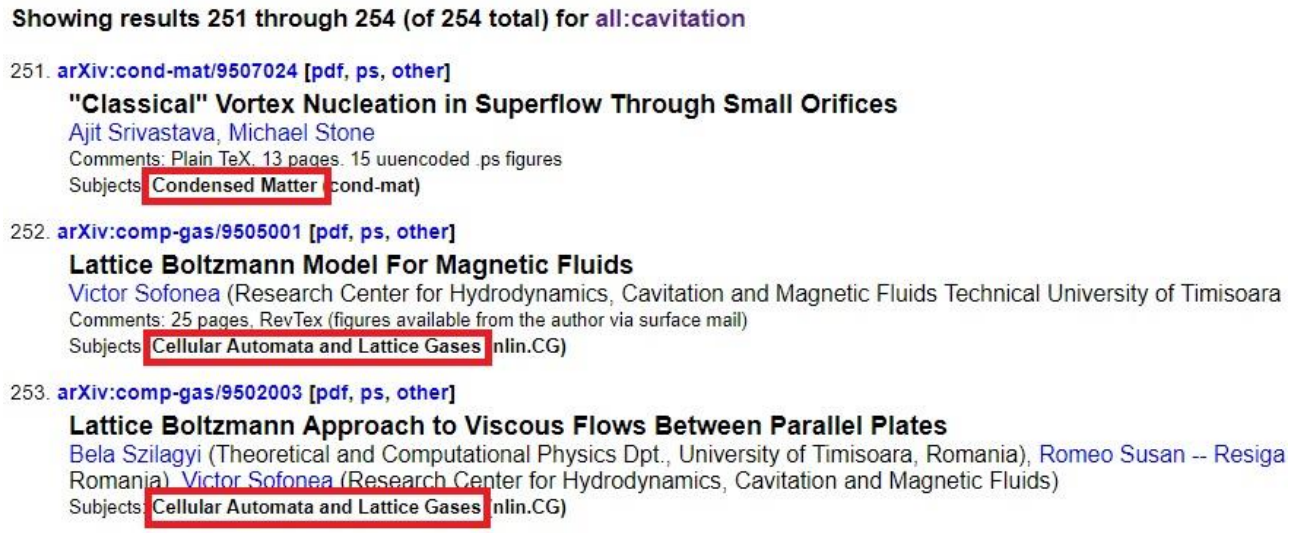


Рисунок 3.13 – Виокремлені назви наукових напрямків, що будуть брати участь у подальшій роботі

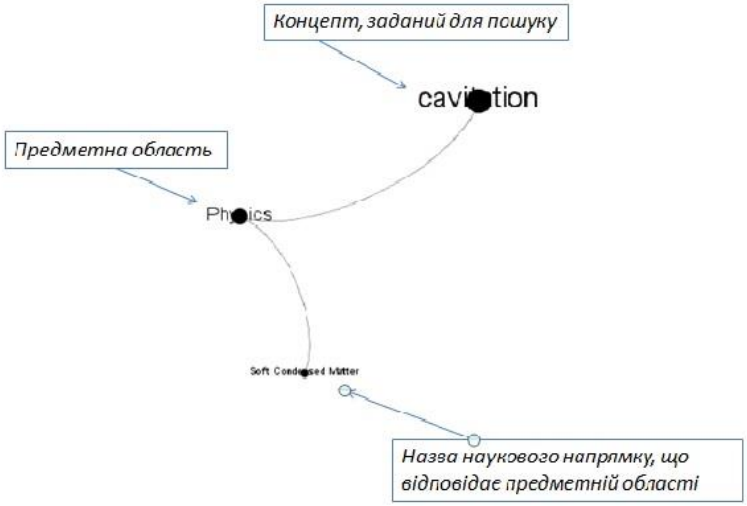


Рисунок 3.14 – Початковий етап побудови мережі предметних областей

Для візуалізації отриманих результатів – отримані дані імпортуються до середовища Gephi.

З огляду на поставлену задачу постає необхідним здійснення оцінки результатів, отрманих в рамках пошуку за заданим концептом - обрахунку парамнтів для заданого концепту.

а) Вага частоти набору документів (collection frequency weight - CFW) [113], для обрахунку якого до уваги беруться такі показники:

- n – кількість документів, в яких було знайдено заданий концепт;
- N – загальна кількість документів, за якими відбувався пошук.

Відповідно, CFW = log N – log n.

б) також для заданого поняття обраховується науковий напрямок, який набільшу кількість разів зустрічається у результатах пошуку і є притаманним для заданого для пошуку концепту.

Показник обраховується за наступною формулою:

$$TF = \frac{k_i}{n}, \tag{3.7}$$

де k_i – це кількість разів, коли назва наукового напрямку зустрічається в переліку результатів пошуку, а n – загальна кількість предметних областей, які було виокремлено для побудови мережі для даного поняття.

Для оцінки текстового пошуку, що було здійснено було обраховано повноту представлення результатів, що може бути порівняна із експертною думкою для визначення коректності здійснення пошуку [114]. Повнота характеризує здатність системи знаходити необхідні користувачеві результати, але не враховує кількість нерелевантних документів, що видаються користувачеві.

Повнота (recall, r) обчислюється як співвідношення знайдених релевантних документів до загальної кількості релевантних документів:

$$r = \frac{a}{a+c}, \tag{3.8}$$

де a – релевантні документи знайдені системою,

c – релевантні документи, що не знайдені системою.

Повнота характеризує здатність системи знаходити потрібні користувачеві документи, але не враховує кількість нерелевантних документів, що надаються користувачеві.

Для реалізації поставленої задачі було застосовано наступні методи:

- Метод аналізу текстових масивів, що застосовується для виокремлення із тексту інформації щодо наукового піднапрмяку, до якого віднесено публікацію, а також виокремлення публікацій, що вже опубліковано в наукових виданнях.
- Методи математичної лінгвістики – формування словників та оцінку текстового пошуку та співставлення.
- Методи статистичного аналізу, що дозволяє провести підрахунки співвідношення кількості публікацій – відповідно піднапрямків і напрямків.
- За теорії графів було побудовано візуалізацію результатів роботи.

Висновки до розділу 3

На сьогодні розроблені моделі, алгоритми та їх програмна реалізація можуть виступати допоміжним інструментарієм в оцінці результативності досліджень, порівняльному аналізі та прийнятті рішень щодо надання підтримки науковим проектам для розвитку за результатами дослідження.

Варто відмітити, що відкритість ресурсів, на базі яких було розрослено і реалізовано зазначені алгоритми дозволить користувачам формувати запити та відповідні масиви інформації для роботи та прийняття рішень.

РОЗДІЛ 4
ПРАКТИЧНА РЕАЛІЗАЦІЯ АЛГОРИТМІВ РОЗШИРЕННЯ МОЖЛИВОСТЕЙ ІНФОРМАЦІЙНИХ СИСТЕМ НА БАЗІ НАУКОМЕТРИЧНОЇ ІНФОРМАЦІЇ

4.1 Побудова мережі предметних областей та мережі співаторів на базі ресурсу Google Scholar Citations за заданими тегами

Запропоновану у роботі модель побудови предметної області та розроблений та реалізований на її базі алгоритм було апробовано для кількох предметних областей [115].

4.1.1 Модель предметної області та мережа співавторів для напрямку фізичної оптики

Для побудови моделі предметної області експертним шляхом було визначено базовий тег ‘physical optics’. Тег було задано англійською, мовою, отже результати, що відображаються були взяті із профілів авторів, що зазначаються відповідну інформацію щодо наукових напрямків англійською мовою.

Шляхом сканування ресурсу Google Scholar Citations було побудовано мережу з наступними параметрами:

- кількість вузлів – 401;
- кількість ребер – 670;
- середня вага вузлів – 1, 67;
- діаметр – 6;
- середній найкоротший шлях – 2, 48.

На рисунку 46 зображено фрагмент мережі предметної області для заданого наукового напрямку.

Візуалізацію результатів побудови мереж було реалізовано засобами Gephi.

Відповідно до запропонованого у розділі алгоритму мережа співавторів була побудована беручи до уваги обмеження до числа вузлів, що сканувалися.

Порядок проходження алгоритму та перелік дослідників та їх теги відображено на рисунку 4.1.

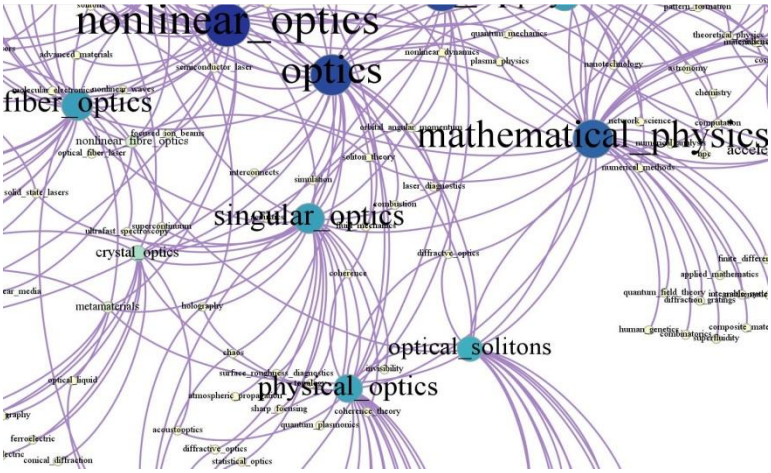


Рисунок 4.1 – Фрагмент мережі предметної області для напрямку фізична оптика

В результаті зондування наукометричного ресурсу було отримано мережу співавторів з наступними параметрами:

- кількість вузлів – 207;
- кількість ребер – 384;
- середня вага вузла – 2, 34;
- діаметр – 7;
- середній найкоротший шлях – 4, 23.

Tiberiu Tudor physical_optics polarization coherence lasers quantum_optics
Sabino Chavez-Cerda optics mathematical_physics physical_optics diffractve_optics optical_solitons
David Snchez-de-la-Llave optics physical_optics fourier_optics_and_signal_processing holography
Miguel A. Bandres physics optics photonics
Johannes Courtial physics optics ray_optics holography
Mark R Dennis mathematical_physics optics singular_optics topology
Franco Nori condensed_matter_physics quantum_optics quantum_information physics superconductivity
Gran Johansson quantum_physics quantum_computing microwave_quantum_optics the_dynamical_casimir_effect mesoscopic_superconductivity
Abraham G. Kofman quantum_physics quantum_information quantum_optics laser_physics solid_state_qubits
Skab Ihor physical_optics singular_optics crystal_optics piezo_and_electrooptics acoustooptics
Eduard Carcol'' physical_optics seismology computers
Neill Lambert physics quantum_optics quantum_computing nano_mechanics quantum_mechanics
Arend G. Dijkstra theoretical_chemical_physics nonlinear_optics open_quantum_systems
B. M. Rodrguez-Lara quantum_optics optical_physics
Suren A. Chilingaryan quantum_optics_and_quantum_information quantum_physics quantum_mechanics
Myun-Sik Kim metrology interferometry physical_optics phase_anomaly microlens
G. Rodriguez Zurita physical_optics interferometry fourier_optics
Vlokh Rostyslav physical_optics
Karol Bartkiewicz quantum_physics quantum_optics quantum_information
Anirban Pathak physics quantum_information quantum_optics
Swapan Mandal quantum_optics laser_spectroscopy quantum_information_theory mathematical_physics
Ioannis Besieris stochastic_linear_and_nonlinear_wave_propagation phase_space_techniques wave_localization

Рисунок 4.2 – Порядок роботи алгоритму – автори та визначені ними теги

Використовуючи кластерний аналіз можна визначити найближчі групи дослідників – співавторів, наукові школи та можливі експертні групи (рисунок 4.3).

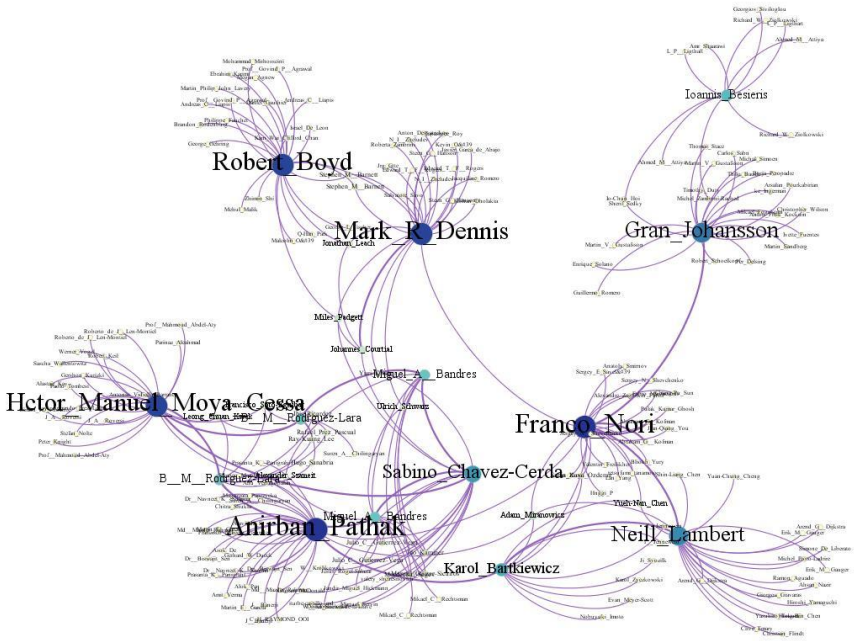


Рисунок 4.3 – Візуалізація мережі співавторів за напрямком фізична оптика

Якщо залишити тільки структурно вагомі вузли і ребра, з використанням Gephi, можлива кластеризація початкової мережі, та більш об’єднаних підгруп дослідників (рисунок 4.4).

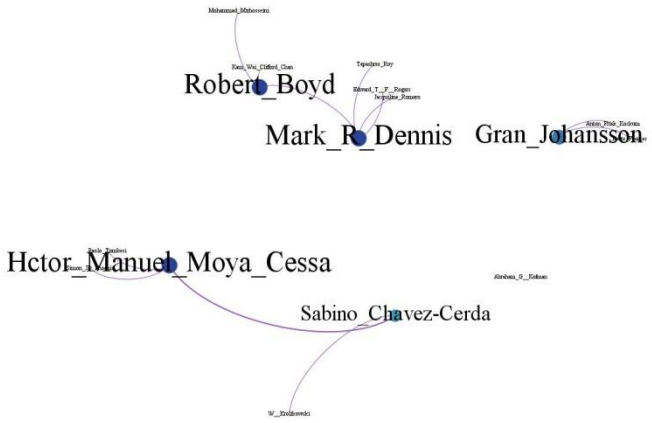


Рисунок 4.4 – Найбільші кластери представленої мережі

Розроблені інтрументи є актуальним додатком для наукометричних баз даних, що дозволять значно розширити існуючі можливості та представляє велику кількість аналітичної інформації не тільки для дослідників, але й для наукових інститутцій, як спосіб моніторингу динаміки активності вчених та їх кооперації, та

також можуть бути рекомендацією для формування політики грантової підтримки країни.

4.1.2 Модель предметної області та мережа співавторів для обмеженого набору тегів

Реалізацію запропоновано підходу до побудови мережі предметної області та мережі співавторів було реалізовано для 6 визначених тегів, заданих англійською мовою: computer, networks, language, information, complex, text. При цьому було обмежено кількість вузлів, що скануються до 1000.Фрагмент мережі співавторів представлено на рисунку 4.5.

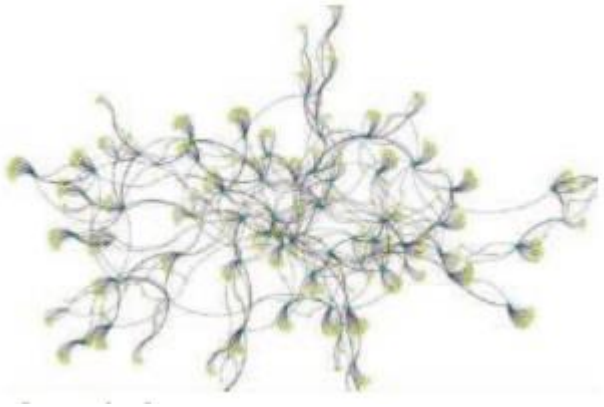


Рисунок 4.5 – Фрагмент мережі співавторів, що побудована за заданими тегами.

Динамка зросту мережі має лінійний тренд через широку предметну область (рисунок 4.6).

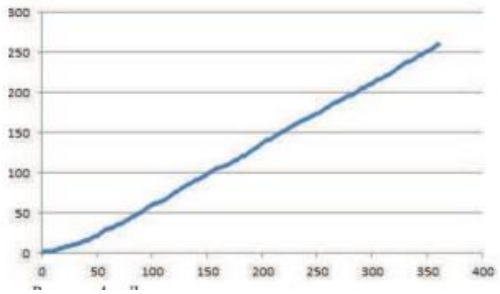


Рисунок 4.6 – Динаміка росту мережі співавторів, що побудована на базі заданих тегів (горизонтальна вісь – кроки вибору вузлів(співавторів), вертикальна – кількість вузлів.

4.2 Онтологія понять на базі енциклопедичного ресурсу Wikipedia

Апробацію алгоритму щодо побудови онтології для заданого поняття ба заі ресурсу Wikipedia, було здійснено для кількох понять:

- поняття, що виражене одним словом – Scientometrics.
- поняття, що виражене словосполученням – mission statement.
- поняття, що виражене у вигляді власного імені – Evgeny Paton.

4.2.1 Поняття ‘Scientometrics’

З огляду на те, що в рамках роботи було проведено аналіз ресурсів, що містять наукометричну інформацію, а також запропоновано нові наукометричні показники, апробацію алгоритму було виконано для поняття ‘Scientometrics’ [20].

Важливим аспектом при проведенні та перевірці результатів було викоремити результати, що стосуються безпосередньо поняття ‘Scientometrics’ та не враховують результати, що стосуються наукового журналу Scientometrics інформація про який графічно відображається в статтях Wikipedia в один і той самий спосіб.

Мережа, побудована за результатами має наступні характеристики:

- вузів – 65;
- ребер – 72;

За рахунок розгляду результатів візуалізації мережі можна виокремити основні поняття – заголовки статей Wikipedia, що є найбільш пов’язаними з поняттям Scientometrics:

- Bibliometric;
- Page rank;
- Web of Science;
- Journal of Infometrics;
- Academic Rankings of World Universities.

Візуалізацію мережі представлено на рисунку 4.7.

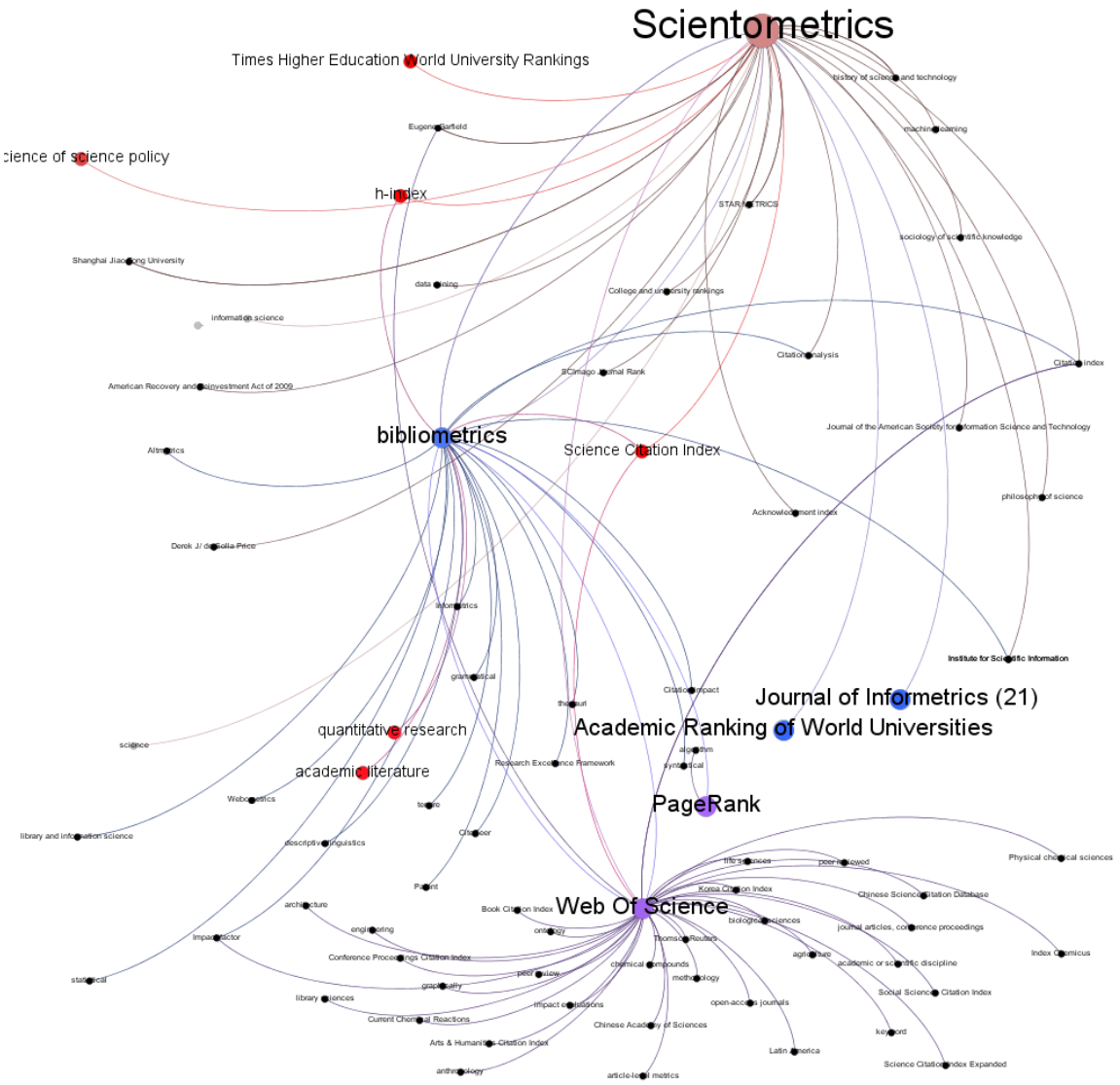


Рисунок 4.7 – Візуалізація онтології побудованої для поняття ‘Scientometrics’

4.2.2 Поняття ‘mission statement’

Метою побудови онтології для поняття ‘mission statement’ було формування низки понять, що корелюються із заданим, для розвитку та вдосконалення визначення міссії організації – задля врахування всіх можливих аспектів, що стосуються даного поняття.

За результатами було побудовано мережу, що містить:

- вузлів – 40;

- ребер – 49.

Базова сторінка, що відображає інформацію про поняття містить – 21 гіперпосилання. Із 21 сторінки, що було просканковано за гіперпосиланнями лише 4 містили інформацію про заданий концепт.

В середовищі Gephi мережі було візуалізовано і отримано результаті зображені на рисунку 4.8.

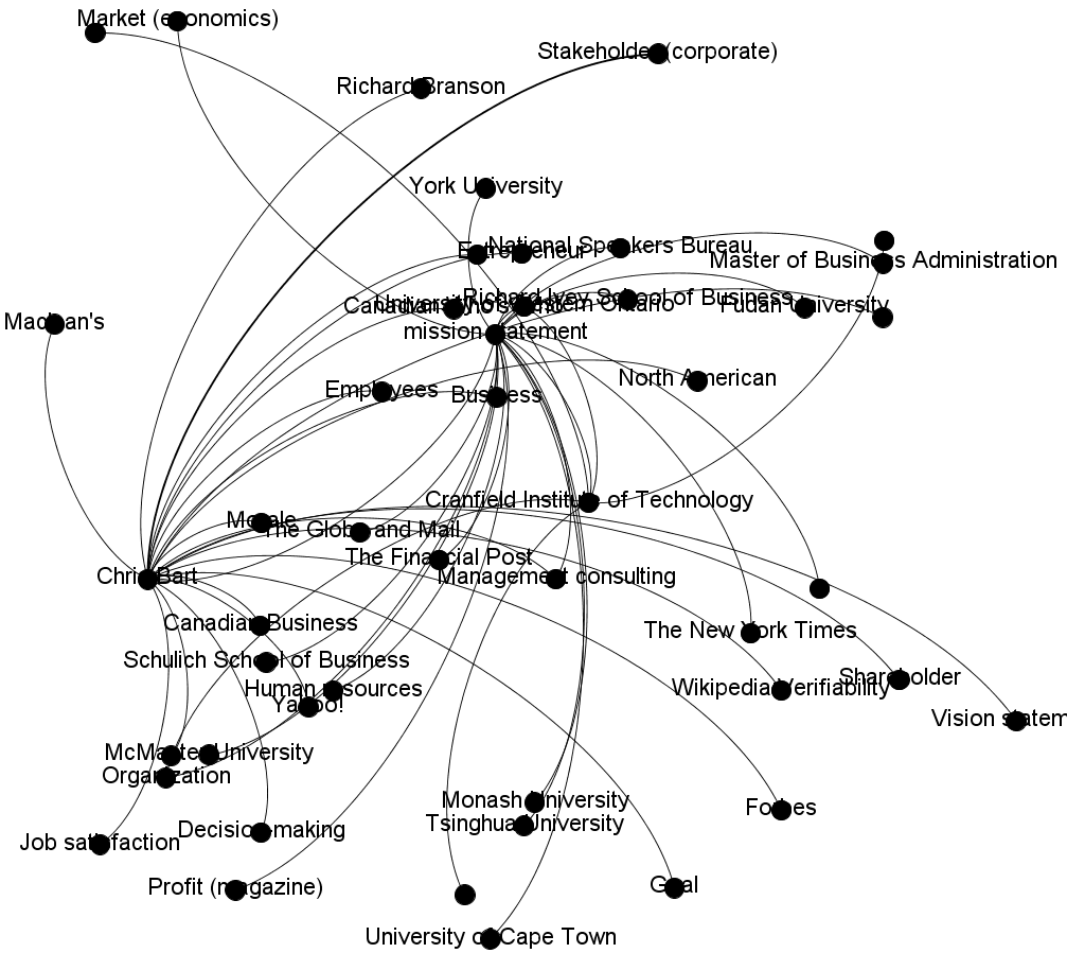


Рисунок 4.8 – Візуалізація онтології для поняття ‘mission statement’

4.2.3 Побудова онтології для імені Evgeny Paton

Важливим етапом апробації запропонованих моделей для побудови онтології за даними енциклопедичного ресурсу Wikipedia є вживання в рамках поняття для пошуку – власного імені.

Для даного прикладу було обрано ім’я - Evgeny Paton.

Мережа складається з:

- 77 вузлів;
- 87 ребер.

Візуалізація онтології представлена на рисунку 54.

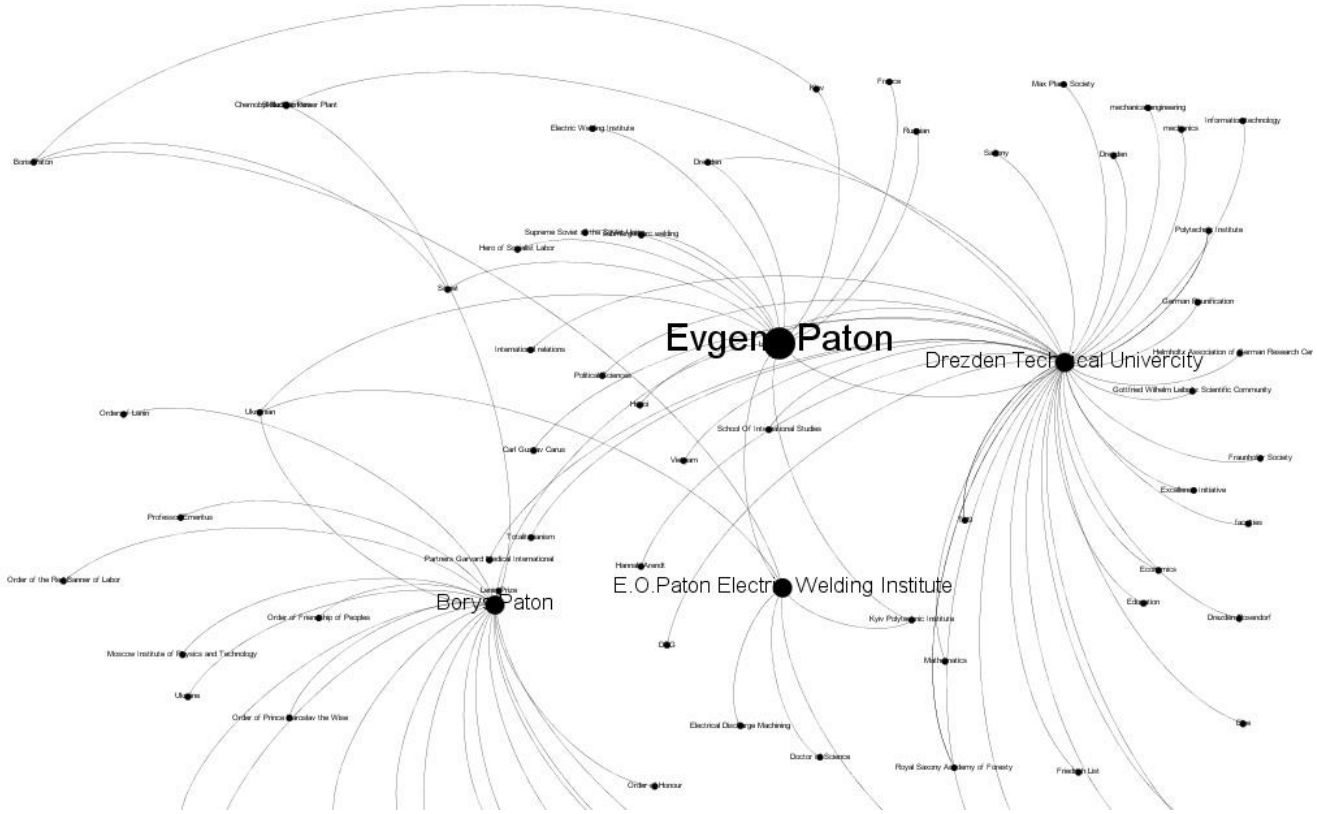


Рисунок 4.9 – Онтологія для імені Evgeny Paton

4.3 Вікі-індекс популярності автора он-лан енциклопедії Вікіпедія

Наведені у розділі 3 алгоритми було реалізовано у вигляді програмного комплексу, за допомогою якого будуються моделі предметних областей та Вікі-індекси популярності авторів [116].

Як приклад розрахунку вікі-індекса зазначимо наступний: припустимо, стаття з Вікіпедія з максимальною кількістю бібліографічних посилань на автора Дж. Сміта (в заданій предметній області) містить 100 посилань. Другий за посиланнями – 20 документів, третій – 10, четвертий – 5, п’ятий – 5, і наступні

чотири – по 1 посиланню. Тобто маємо ряд із наступних значень: R_1=100, R_2=20, R_3=10, R_4=5, R_5=5, R_6=1, R_7=1, R_8=1, R_9=1

1 стаття містить кількість посилань не менше ніж R_1=100;

2 статті містить кількість посилань не менше ніж R_2=20;

3 статті містить кількість посилань не менше ніж R_3=10;

4 статті містить кількість посилань не менше ніж R_3=4.

5 статей містять кількість посилань не менше ніж R_5=5.

Не існує 6 статей, що містять кількість посилань не менше ніж 6.

В даному випадку:

$N = 9,\ \ WH = 5,$

Таким чином, $WI = 5 \times \sqrt{9} = 15.$

Наведемо приклади розрахунку Вікі-індексів для трьох авторів: Альберт Ейнштейн, Енріко Фермі, Бенуа Мандельброт.

На рисунку 4.10 наведено фрагменти трас виконання програми зондування Wikipedia, на яких відображаються поняття, до яких відбувається перехід від вихідних понять, до понять, що містять ім’я автора та перевірочне слово.

Albert_Einstein

1: Albert_Einstein

SCI Links (1): 174
0 Rd +: Ulm
1 Rd -: German_Empire
2 Rd -: Statelessness
3 Rd -: Switzerland
4 Rd -: Kingdom_of_Prussia
5 Rd -: Free_State_of_Prussia
6 Rd -: Weimar_Republic
7 Rd -: Physics
8 Rd -: Philosophy
9 Rd -: Swiss_Patent_Office
10 Rd -: Bern
11 Rd -: University_of_Bern
12 Rd -: University_of_Zurich
13 Rd +: ETH_Zurich
14 Rd -: Kaiser_Wilhelm_Institute
15 Rd -: German_Physical_Society
16 Rd +: Leiden_University

Enrico_Fermi

1: Enrico_Fermi

SCI Links (1): 28
0 Rd -: Rome
1 Rd -: Chicago
2 Rd -: Physics
3 Rd +: Leiden_University
4 Rd -: University_of_Florence
5 Rd -: Columbia_University
6 Rd +: University_of_Chicago
7 Rd -: Alma_mater
8 Rd -: Doctoral_advisor
9 Rd +: Luigi_Puccianti
10 Rd +: Max_Born
11 Rd -: Paul_Ehrenfest
12 Rd +: Harold_Agnew
13 Rd -: Edoardo_Amaldi
14 Rd +: Owen_Chamberlain
15 Rd +: Geoffrey_Chew
16 Rd +: Jerome_Isaac_Friedman

Benoit_Mandelbrot

1: Benoit_Mandelbrot

SCI Links (1): 47
0 Rd +: Mandelbrot_set
1 Rd -: Warsaw
2 Rd -: Second_Polish_Republic
3 Rd -: Mathematics
4 Rd -: Aerodynamics
5 Rd -: Yale_University
6 Rd -: IBM
7 Rd -: Alma_mater
8 Rd -: University_of_Paris
9 Rd -: Eugene_Fama
10 Rd +: Ken_Musgrave
11 Rd -: Murad_Taqqu
12 Rd +: Mandelbrot_set
13 Rd +: Chaos_theory
14 Rd +: Fractal
15 Rd -: Johannes_Kepler
16 Rd +: Szolem_Mandelbrojt

Рисунок 4.10 – Фрагменти трас програми зондування Wikipedia

На рисунку 4.11, аркуш 131-132 наведено візуалізацію за допомогою програми Gephi фрагментів моделей предметних областей, що було отримано шляхом зондування Вікіпедія. Параметри отриманих мереж (моделей предметних областей), вузлами яких вступають поняття з Wikipedia, наступні, для мережі, що відповідає моделі предметної області авторів:

- Альберт Ейнштейн: вузлів 718, ребер: 22111, середній степінь вузла: 62, діаметр графа: 4, середній коефіцієнт кластеризації: 0,26, найбільші вузли (таблиця 4.1, рисунок 4.11, аркуш 131) :

Таблиця 4.1 – Найбільші вузли мережі для заданого автора - Альберт Ейнштейн.

Поняття	Степінь вузла
Quantum_nonlocality	188
Alain_Aspect	181
Hermann_Weyl	177
Paul_Dirac	174
Electromagnetic_radiation	174
Isaac_Newton	169
Galileo_Galilei	169
Wolfgang_Pauli	169
General_relativity	167

- Енріко Фермі: вузлів 605, ребер: 22079, середній степінь вузла: 73, діаметр графа: 4, середній коефіцієнт кластеризації: 0,47, найбільші вузли (таблиця 4.2, рисунок 4.11, аркуш 132):

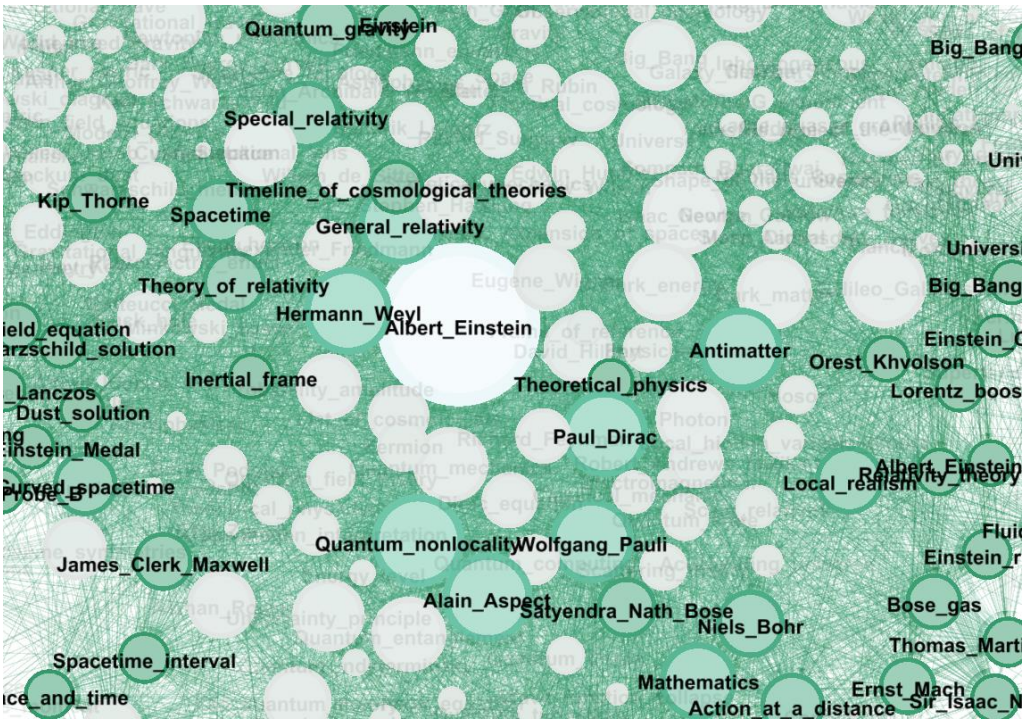
Таблиця 4.2 – Найбільші вузли мережі для заданого автора – Енріко Фермі

Поняття	Степінь вузла
Enrico_Fermi	440
Nobelium	206
Transuranic_element	206
Particle_physics	204
Mendelevium	204
Einsteinium	204
Berkelium	203
Radioactive_decay	195
Radioactive	190
Particle_accelerators	188

- Бенуа Мандельброт: вузлів 34, ребер: 259, середній степінь вузла: 15,2, діаметр графа: 3, середній коефіцієнт кластеризації: 0,52, найбільші вузли (таблиця 4.3, рисунок 4.11, аркуш 132):

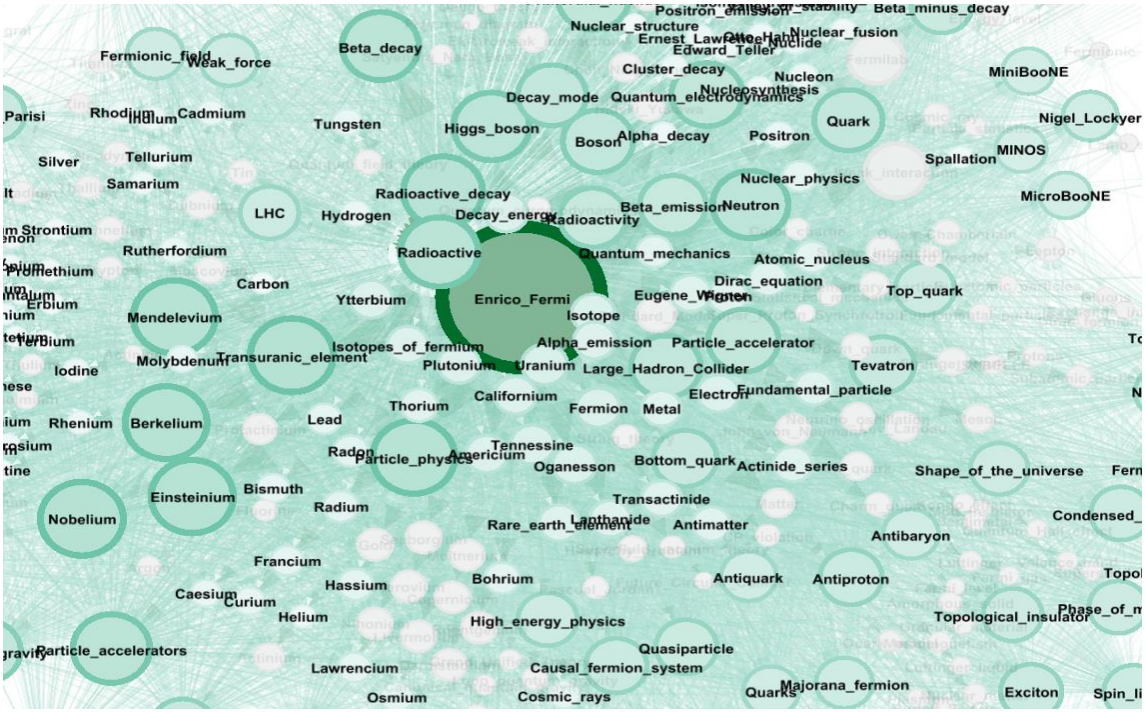
Таблиця 4.3 – Найбільші вузли мережі для заданого автора - Бенуа Мандельброт

Поняття	Степінь вузла
Benoit_Mandelbrot	22
Pattern	20
Chaos_theory	18
Patterns_in_nature	18
Hausdorff_dimension	17
Patterns	16
Fractal	15
Fractal_dimension	15
Fractal_geometry	15
Fractals	15

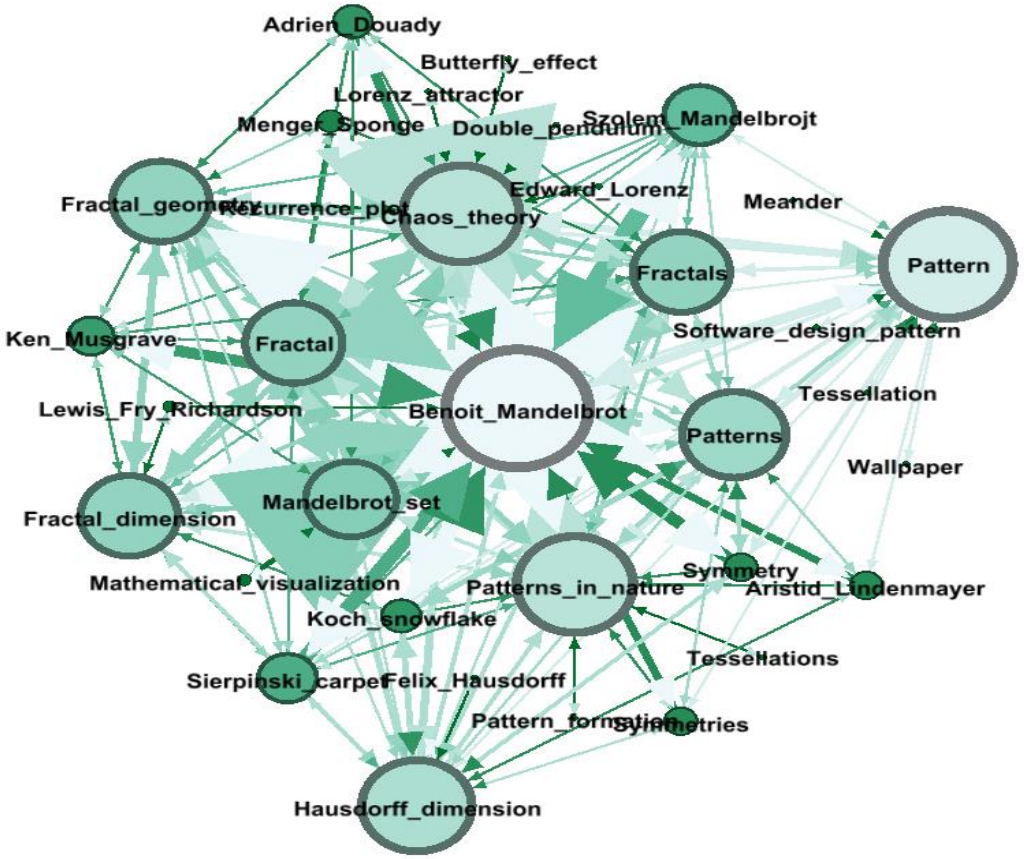


а)

Рисунок 4.11 – Фрагменти моделей предметної області: а) - Альберт Ейнштейн; б) – Енріко Фермі; в) – Бенуа Мандельброт


Positron_emission
Beta_minus_decay
Nuclear_structure
Nuclear_fusion
Ernest_Lawrence
Edward_Teller
Nuclide
Fermionic_field
Weak_force
Beta_decay
Cluster_decay
Nucleon
MiniBooNE
Decay_mode
Quantum_electrodynamics
Nucleosynthesis
Parisi
Rhodium
Cadmium
Helium
Tungsten
Higgs_boson
Quark
Nigel_Lockyer
Boson
Alpha_decay
Positron
Silver
Tellurium
Spallation
MINOS
Samarium
Nuclear_physics
Radioactive_decay
MicroBooNE
LHC
Hydrogen
Decay_energy
Beta_emission
Neutron
Radioactivity
Strontium
Rutherfordium
Radioactive
Quantum_mechanics
Atomic_nucleus
Promethium
Carbon
Dirac_equation
Ytterbium
Enrico_Fermi
Eugene_Wigner
Top_quark
Erbium
Mendelevium
Isotope
Terbium
Isotopes_of_fermium
Alpha_emission
Particle_accelerator
Molybdenum
Transuranic_element
Plutonium
Uranium
Large_Hadron_Collider
Tevatron
Iodine
Californium
Electron
Fundamental_particle
Thorium
Fermion
Metal
Rhenium
Berkelium
Lead
Tennessine
Radon
Americium
Particle_physics
Oganesson
Bottom_quark
Actinide_series
Shape_of_the_universe
Einsteinium
Bismuth
Transactinide
Radium
Nobelium
Lanthanide
Antimatter
Rare_earth_element
Antibaryon
Condensed_
Francium
Hassium
Bohrium
Antiquark
Antiproton
Caesium
Curium
Helium
High_energy_physics
Topological_insulator
Particle_accelerators
Quasiparticle
Lawrencium
Causal_fermion_system
Osmium
Cosmic_rays
Quarks
Majorana_fermion
Exciton

б)


Adrien_Douady
Butterfly_effect
Lorenz_attractor
Menger_sponge
Double_pendulum
Szolem_Mandelbrojt
Fractal_geometry
Recurrence_plot
Chaos_theory
Edward_Lorenz
Meander
Fractals
Pattern
Ken_Musgrave
Fractal
Software_design_pattern
Lewis_Fry_Richardson
Benoit_Mandelbrot
Tessellation
Patterns
Wallpaper
Mandelbrot_set
Fractal_dimension
Mathematical_visualization
Patterns_in_nature
Symmetry
Aristid_Lindenmayer
Koch_snowflake
Tessellations
Sierpinski_carpet
Felix_Hausdorff
Pattern_formation
Symmetries
Hausdorff_dimension

в)

Принцип побудови Вікі-індекса автора відрязняється від тих індексів, що використовуються на сьогодні в наукометрії в першу чергу врахуванням цитувань не тільки з науковий публікацій, а й з популярних сторінок самого сервісу Вікіпедія. Таким чином, в принципі, можна отримати індекс популярності автора в межах даного сервісу. Це є істотним, адже Вікіпедія на сьогодні є найбільшою і найпопулярнішою серед користувачів інтернету он-лайн енциклопедією.

В роботі запропоновано технологію «швидкого» розрахунку Вікі-індексу автора, яка дозволяє реалізувати розрахунок у вигляді окремого сервісу.

Можна зауважити, що ресурс Вікіпедія, як і ресурс Google Scholar Citations, що розглядалися раніше, є зручною щодо доступу до інформації, не передбачає створення власного профілю користувача для доступу до інформації, доступ є необмежений.

4.4 Побудова мережі предметних областей на базі ресурсу препринтів arXiv

Розроблений на базі запропонованої у розділі 3 моделі алгоритм було апробовано на кількох концептах [117], [118].

4.4.1 Концепт “cavitation”

В результаті роботи було отримано наступні результати [112].

а) за результатами пошуку для заданого концепту було системою знайдено 254 публікації.

б) за рахунок реалізації алгоритму було виокремлено 5 наукових напрямків за якими було розміщенно публікації на ресурсі, це:

- Physics – 25 підгруп за напрямком;
- Computer Science – 5 підгруп за напрямком;
- Mathematics – 4 підгрупи за напрямком;
- Quantitative Biology – 2 підгрупи;
- Statistics – тільки 1 підгрупа.

Отже за результатами можна впевнено сказати, що задане поняття є притаманним для 5 предметних областей і серед публікацій, що розміщені на ресурсі, найбільше – з напрямку – фізика, зокрема – фізика рідин (Fluid Physics) – 82 публікації.

За результатами дослідження було побудовано мережу предметних областей для заданого концепту, і в середовищі Gephi виконано візуалізацію. На рисунку 4.12 показано візуалізацію отриманих результатів.

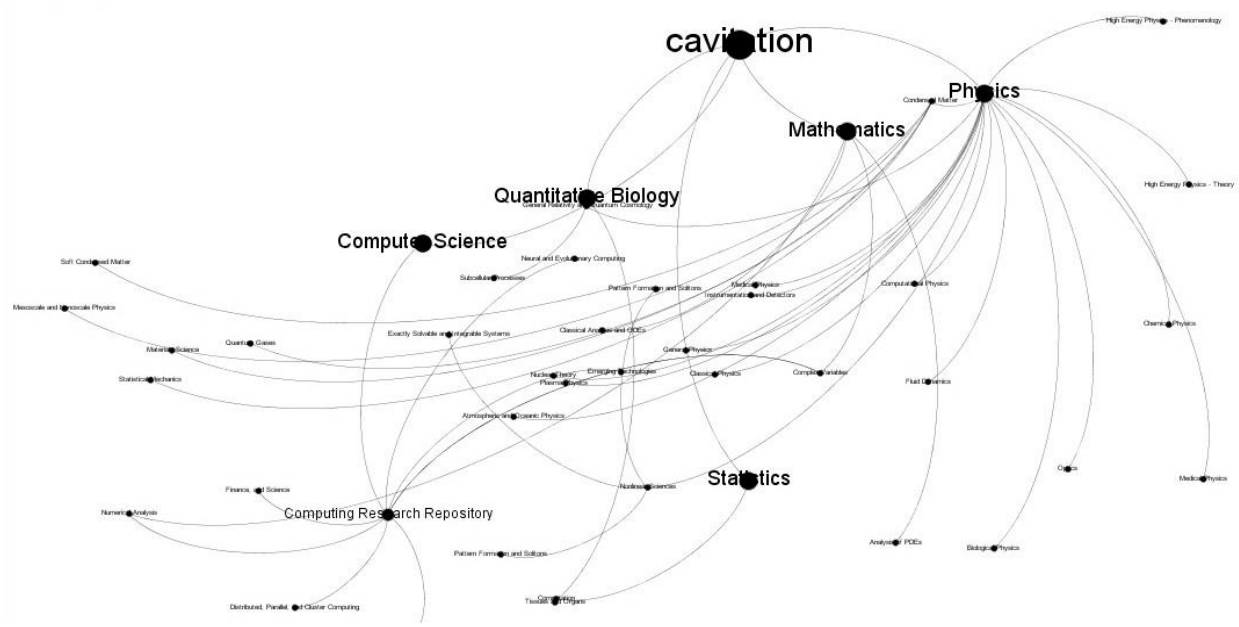


Рисунок 4.12 – Мережа предметних областей для поняття «cavitation»

Вага частоти набору документів (collection frequency weight - CFW) для заданого концепту буде дорівнювати:

$\mathrm{CFW} = \log N - \log n = 3.734$, де загальна кількість документів ресурсу становить -

N – загальна кількість публікацій на ресурсі, цей показник становить 1,377,332 (дані на 23.04.2018 року);

n – кількість публікацій, що містять заданий для пошуку концепт, і дорівнює 254.

Для даної предметної області показник Term frequency буде обрахований наступним чином:

$$TF = \frac{25}{5} = 5$$

Для оцінки текстового пошуку, що було здійснено було обраховано повноту представлення результатів, що можуть бути порівняні із експертною думкою для визначення коректності здійснення пошуку.

Повнота (для кожного виокремленого наукового напрямку визначеного системою):

Physics, $r = \frac{25}{25+229} = 0{,}09$

Computer Science, $r = \frac{5}{254} = 0{,}01$

Mathematics, $r = \frac{4}{254} = 0{,}02$

Quantitative Biology, $r = \frac{2}{254} = 0{,}008$

Statistics, $r = \frac{1}{254} = 0{,}004$

4.4.2 Концепт ‘Ukraine’

Задачу було також вирішено для концепту «Ukraine», що дозволило отримати результати щодо кількості публікацій афілійованих українськими інституціями, були представлені на наукових заходах в Україні, а також оцінити відсоток препринтів, що в подальшому були опубліковані [112].

Алгоритм було реалізовано мовою програмування Java в середовищі Eclipse.

За заданим концептом було виокремлено 619 публікацій. До уваги було взято 200 публікацій за період – 2018 – 2009 роки. Такий вибір щодо обсягу публікацій зумовлений актуальністю і можливістю перевірки працездатності системи – реалізації поставленої задачі.

За результатами роботи розробленої системи було отримано нові масиви інформації щодо розподілу публікацій між науковими напрямками та відповідними піднапрямками на ресурсі препринтів arXiv. За результатами отриманої інформації в середовищі Gephi було візуалізовано результати у вигляді ненаправленого графу.

Також окремий пошук було запроваджено для виокремлення переліку публікацій, розміщених на ресурсі, що вже було опубліковано, і також візуалізовано в середовищі Gephi.

Результатом роботи системи є отримання наступних даних щодо кількості публікацій за визначеними ресурсом arXiv науковими напрямами і піднапрямами:

- Physics – 111 публікацій за 32 піднапрямками.
- Mathematics – 52 публікації за 13 піднапрямками.
- Computer Science – 28 публікацій за 13 піднапрямками.
- Quantitative Finance – 4 публікації за 2 піднапрямками.
- Quantitative Biology – 3 публікації за 2 піднапрямками.
- Statistics – 2 публікації за 2 піднапрямками.

Жодних публікацій за заданим концептом не містили напрямки – Electrical Engineering and System Science та Economics.

Серед вищезазначеного масиву публікацій розміщені у наукових виданнях або опубліковані в матеріалах конференцій або тези, наступні:

- Physics – 55 публікацій за 25 піднапрямками.
- Mathematics – 14 публікацій за 9 піднапрямками.
- Computer Science – 14 публікацій за 8 піднапрямками.
- Quantitative Biology – 1 публікація за 1 піднапрямком.

Вищезазначені результати було візуалізовано у вигляді ненаправленого графа (рисунки 4.13, 4.14). Вершинами графа є наукові напрями і піднапрямки із зазначенням кількості публікацій за заданим концептом, розміщених на ресурсі препринтів.

Графічне представлення інформації є зручним інструментом не тільки для її кращого сприйняття, але й для подальшого використання результатів.

Розмір вершин графа залежить від кількості публікацій, що відповідають науковому напрямку – таким чином одразу можна визначити найбільш популярні серед авторів наукові напрями для розміщення публікацій на ресурсі.

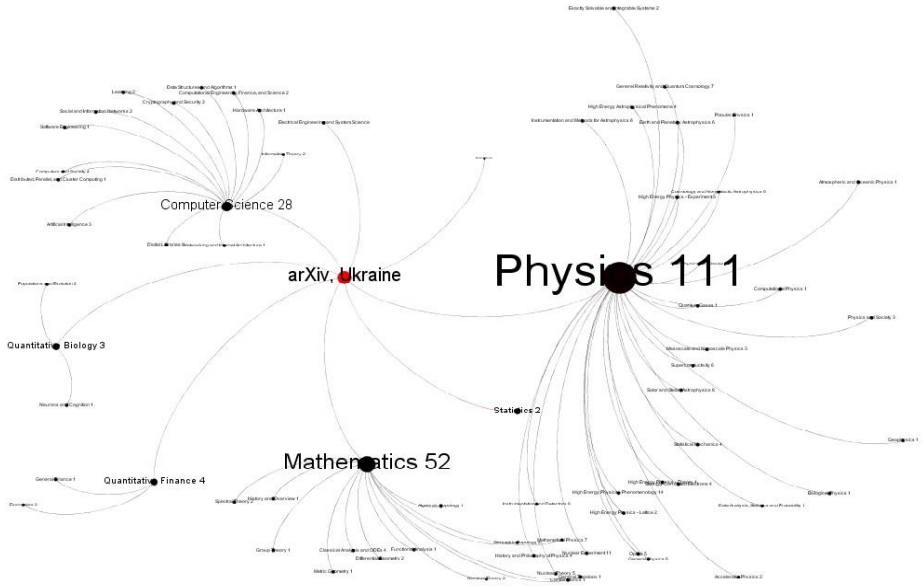


Рисунок 4.13 – Візуалізація результатів пошуку публікацій за заданим концептом на ресурсі arXiv

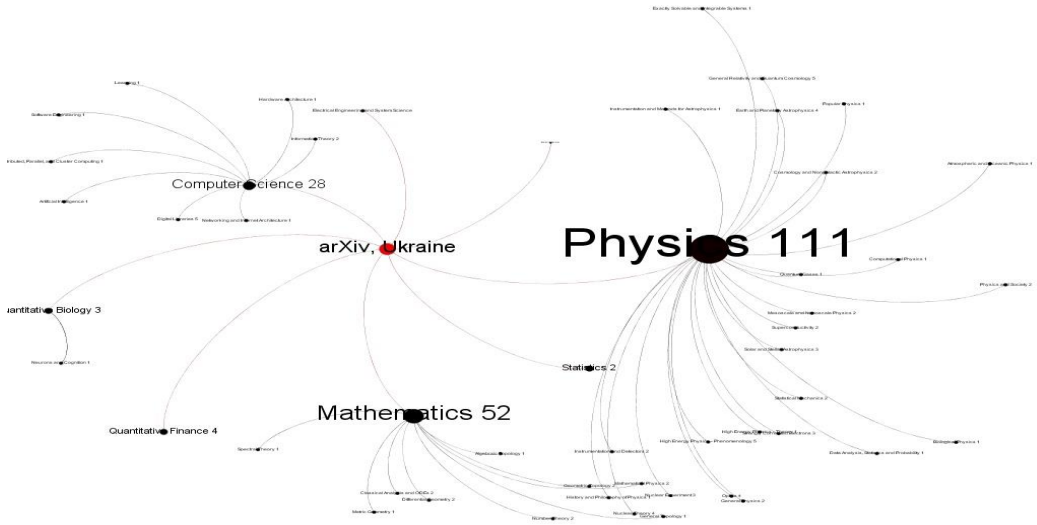


Рисунок 4.14 – Візуалізація результатів пошуку публікацій за заданим концептом, що були опубліковані в наукових журналах та матеріалах конференцій

Візуалізація результатів пошуку препринтів та окремою схемою – тих препринтів, що були опубліковані дає можливість оцінити, яка частка розміщених матеріалів в подальшому були опубліковані.

Відповідно до отриманих результатів можна стверджувати, що робота системи, що виконує пошук результатів надає відносно повну картину щодо представлення публікацій у відповідних наукових напрямках визначених ресурсом.

Практична реалізація результатів дослідження дозволяє припустити, що поставленої мети було досягнуто. За рахунок розробки алгоритмів для запропонованих моделей, що дозволяють сформувати нові масиви інформації на базі існуючих ресурсів наукової інформації.

Висновки до розділу 4

Практичну реалізацію було реалізовано на базі розрослених алгоритмів засобами мови програмування Java в середовищі Eclipse. Візуалізацію результатів було представоено в середовищі Gephi.

Всі результати і покрокове виконання алгоритму було перевірено за рахунок використання понять, результатами обробки яких є невеликі обсяги даних.

В той же час одним із варіантів розвитку дослідження – є реалізація запропонованих алгоритмів мовою програмування Python, що передбачає роботу із великими масивами інформації і що дозвилить обробку великих наборів даних.

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 мультидисціиплінарних 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 наукові 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 реалізаці 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 данного 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'зяні 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 предмтених областях. Реалізовані моделі та програмна реалізація алгоритмів дозволить здійснювати пошук міждисциплінарної складової для пошуку партнерів та формування успішних колаборацій для реалізації складних міждисциплінарних проектів 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 рботи.

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 ВИКОРИСТАНИХ ДЖЕРЕЛ

1. What does the General Data Protection Regulation Govern? European Commission [Internet], [updated 2018, June, cited 2018 July 11]. Available from: https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-does-general-data-protection-regulation-gdpr-govern_en
2. Jeremy B. White, Twitter announces new privacy policy ahead of European data law, Independent Magazine [Internet], [cited 2018 June 16]. Available from: https://www.independent.co.uk/life-style/gadgets-and-tech/news/twitter-privacy-policy-rules-update-personal-data-a8320666.html
3. Katie Collins, Google makes privacy policy clearer than ever to comply with EU law, CNET [Internet], [cited 2018 May 15]. Available from: https://www.cnet.com/news/google-makes-privacy-policy-clearer-than-ever-to-comply-with-eu-gdpr-law/
4. Положення про набори даних, які підлягають оприлюдненню у формі відкритих даних, Урядовий портал [Інтернет], [Цитовано 2017, 11 Гру.]. Доступно на: https://www.kmu.gov.ua/ua/npas/248573101.
5. Open Access, Wikipedia [Internet], [cited 2018 Mar 16]. Available from: https://en.wikipedia.org/wiki/Open_access
6. H2020 Programme Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020, European Commission Directorate General for Research and Innovation [Internet], [cited 2018 June 16]. Available from: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
7. Regulation Of The European Parliament And Of The Council – establishing Horizon Europe – the Framework Programme for Research and Innovation, laying down its rules for participation and dissemination [Internet], [cited 2018 June 17]. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1540387631519&uri=CELEX%3A52018PC0435

8. 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.

9. Закон України про національну програму інформатизації. Законодавство України. [Інтернет], [Цитовано 2018, 15 Бер.]. Доступно на: https://zakon.rada.gov.ua/laws/show/74/98-%D0%B2%D1%80

10. Hassan NR, Loebbecke C. Engaging scientometrics in information systems. Journal of Information Technology. 2017 Apr 24; 111 (33): 1875-1878.

11. Ranking Web of Universities [Internet], [cited 2018 June 17]. Available from: http://www.webometrics.info/en

12. Falagas ME, Pitsouni EI, Malietzis GA, Pappas G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. The FASEB Journal [Internet]. 2007 Mar [cited 2018 July 12]; 22(2): 112-120/ Available from: https://www.fasebj.org/doi/pdf/10.1096/fj.07-9492LSF

13. Jacso P. Google Scholar: the pros and the cons. Online Information Review.2005: 29(2):208-214.

14. Bar-Ilan J. Which h-index? — A comparison of WoS, Scopus and Google Scholar. Scientometrics. 2008 Feb; 74(2):257-271.

15. Harzing A-WK, W R. Google Scholar as a new source for citation analysis. Ethics in Science and Environmental Politics [Internet]. 2008 June 3 [cited 2017 Jan 25]; 8(1):61-73.

16. Kulkarni AV, Aziz B, Shams I, Busse JW. Comparisons of Citations in Web of Science, Scopus, and Google Scholar for Articles Published in General Medical Journals. JAMA [Internet]. 2009 Sept 9 [cited 2017 Jan 25]; 302(10):1092-1096. Available from: https://jamanetwork.com/journals/jama/fullarticle/184519.

17. Bould MD. References that anyone can edit: review of Wikipedia citations in peer reviewed health science literature. The BMJ [Internet]. 2014 6 March [cited 2017 Jan 25]; 308: g1585. Available from: https://www.bmj.com/content/348/bmj.g1585

18. Kim PT. The Visibility of Wikipedia in Scholarly Publications. First Monday [Internet]. 2011 Aug 1[cited 2017 Jan 25]; 16(8):1-16. Available from:

https://scholarworks.iu.edu/dspace/bitstream/handle/2022/21757/The%20Visibility%20of%20Wikipedia%20in%20Scholarly%20Publications.pdf?sequence=1&isAllowed=y

19. Evans P, Krauthammer M. Exploring the Use of Social Media to Measure Journal Article Impact. AMIA Annual Symposium Proceedings [Internet]. 2011 Oct 22. 2011 [cited 2017 Jan 25]; p.374-381. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243242/
20. Barrena A, Soroa A, Agirre E. Combining Mention Context and Hyperlinks from Wikipedia for Named Entity Disambiguation. ACL Antology [Internet]. 2015 June 4 [cited 2017 Jan 25]; p.101-105. Available from: http://www.aclweb.org/anthology/S15-1011
21. Operational Database Management System. ODBMS.org [Internet]. 2017 June [cited 2017 Jan 25]. Available from: http://www.odbms.org/2017/06/text-annotation-tools/
22. Ferragina P, Scaiella U. Fast and Accurate Annotation of Short Texts with Wikipedia Pages. IEEE Software [Internet]. 2012 Feb [cited 2017 Jan 25]; 29(1):70-75. Available from: https://ieeexplore.ieee.org/abstract/document/6035657/
23. Huynh DT, Cao TH, Pham PHT, Hoang TN. Using Hyperlink Texts to Improve Quality of Identifying Document Topics Based on Wikipedia. In: 2009 International Conference on Knowledge and Systems Engineering [Internet]; 2009 Oct 13-17; Hanoi, Vietnam; 2009 Dec [cited 2017 Jan 25]; Available from: https://ieeexplore.ieee.org/abstract/document/5361697/?part=1
24. Gabay D, Ben-Eliahu Z, Elhadad M. Using Wikipedia Links to Construct Word Segmentation Corpora. AI Magazine [Internet]. 2008 [cited 2017 Jan 25]; 15(011): 61-63. Available from: http://www.aaai.org/Papers/Workshops/2008/WS-08-15/WS08-15-011.pdf
25. Hu X, Zhang X, Lu C, Park EK, Zhou X. Exploiting Wikipedia as external knowledge for document clustering. In: The 15th ACM SIGKDD international conference on Knowledge discovery and data mining [Internet]. Proceedings; 2009

June 28 – July 1; [cited 2017 Jan 25]. p. 389-396. Available from: https://dl.acm.org/citation.cfm?id=1557066

26. Pedro VC, Niculescu RS, Lita LV. Okinet: Automatic Extraction of a Medical Ontology From Wikipedia. AI Magazine [Internet]. 2008 [cited 2017 Jan 25]; 15(007): 37-42. Available from: http://www.aaai.org/Papers/Workshops/2008/WS-08-15/WS08-15-007.pdf
27. Ngo Q-H, Doan S, Winiwarter W. Using Wikipedia for extracting hierarchy and building geo-ontology. International Journal of Web Information Systems. 2013 June; 8(4): 401-412.
28. Shibaki Y, Nagata M, Yamamoto K. Constructing Large-Scale Person Ontology from Wikipedia. In: Proceedings of the 2nd Workshop on "Collaboratively Constructed Semantic Resources" [Internet]; 2010, Beijing; 2010 [cited 2017 Jan 25]; 1-9. Available from: http://www.aclweb.org/anthology/W10-3501
29. Syed ZS, Finin T, Joshi A. Wikipedia as an Ontology for Describing Documents. AI Magazine [Internet]. 2008 [cited 2017 Jan 25]; 08(024): 136-144. Available from: http://www.aaai.org/Papers/ICWSM/2008/ICWSM08-024.pdf
30. Tzekou P, Stamou S., Kirtsis N, Zotoz N. Quality Assessment of Wikipedia External Links. Engineering and Informatic Depertment of Patras University [Internet]. 2011[cited 2017 Jan 25]. Available from: http://www.dblab.upatras.gr/download/nlp/NLP-Group-Pubs/11-WEBIST_Wikipedia_External_Links.pdf.
31. Raspberry L. Citing Wikipedia. The BMJ [Internet]. 2014 March 16[cited 2017 Jan 25]; 348: g1819. Available from: https://www.bmj.com/content/348/bmj.g1819.full
32. Fariba T, Jamali H. Why and Where Wikipedia Is Cited in Journal Articles?. Journal of Scientometrics. 2013; 2(3): 231.
33. Nakayama K, Pei M, Erdmann M, Ito M, Shirakawa M, Hara T, Nishio S. Wikipedia Mining Wikipedia as a Corpus for Knowledge Extraction. Special Interest Group On Wikipedia Research [Internet]. 2008 [cited 2017 Jan 25]. Available from: http://sigwp.org/en/images/0/06/Wikimania2008.pdf

34. Haque A, Ginsparg P. Positional effects on citation and readership in arXiv. Journal of the Association for Information Science and Technology [Internet]. 2009 Nov [cited 2017 Jan 25]; 60(11): 2203-2218. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21166
35. Bar-Ilan J. Astrophysics publications on arXiv, Scopus and Mendeley: a case study. Scientometrics. 2014; 100(1): 217-225.
36. Davis P, Fromerth M, Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? Scientometrics. 2007; 71(2).
37. Boldt A, Extending ArXiv.org to Achieve Open Peer Review and Publishing. Journal of Scholarly Publishing [Internet]. 2011 Jan [cited 2017 Jan 25]; 42(2): 238-242. Available from: https://www.utpjournals.press/doi/abs/10.3138/jsp.42.2.238
38. Додонов АГ, Ланде ДВ, Путятин В.Г. Компьютерные информационно-аналитические системы. Толковый словарь. Київ; Наукова думка; 2011. 366 с.
39. Андрущенко ВБ. Інформаційно-аналітична діяльність Державного фонду фундментальних 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.
40. Андрущенко ВБ, Кияк БР. Обґрунтування критеріїв оцінювання фундаментальних наукових досліджень. Наука та наукознавство. Грудень 2015; 4(89):67-72.
41. 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 фундамнтальних досліджень (досвід Державного фонду фундаментальних досілджень України); с. 74-81.
42. Андрущенко ВБ, Кияк БР. Анотований збірник проектів спільного конкурсу ДФФД - БРФФД. Київ: ВД «Академперіодика»; 2017. Частина1, Критерії та показники успішної міжнародної наукової співпраці; с. 5-9.

43. Zgorulko Y, Zagorulko G. Ontology-Based Technology for Development of Intelligent Scientific Internet Resources. In: Intelligent Software Methodologies, Tools and Techniques – SoMeT 2015 [Internet]; 2015. Springer. 2015 Sept 1 [cited 2018 June 16]; p. 227-241. Available from: https://link.springer.com/chapter/10.1007/978-3-319-22689-7_17
44. Berestova TF. Information resource studies as a new direction of scientific research: Formulation of the problem. Scientific and Technical Information Processing. 2015; 42(3): 127-134.
45. Das M, Cui R, Campbell DR. Towards methods for systematic research on big data. In: 2015 IEEE International Conference on Big Data (Big Data) [Internet]; 2015 Oct 29 – Nov 1; Santa Clara, CA, USA: IEEE; Dec 28 2015 [cited 2018 June 11]; Available from: https://ieeexplore.ieee.org/abstract/document/7363989/
46. Sayago S. The construction of qualitative and quantitative data using discourse analysis as a research technique. Quality & Quantity. 2015 March; 49(2): 727-737.
47. Huysman **S**, Sala S, Mancini L, Ardente F, Alvarenga RAF, De Meester S, at al. Toward a systematized framework for resource efficiency indicators. Recources, Concervation and Recycling. 2015 Feb; 95: 68-76.
48. Power DJ. Creating a Data-Driven Global Society. Reshaping Society through Analytics, Collaboration, and Decision Support. 2015; 18: 13-28.
49. Савченко ЗВ. *Formation and use of electronic information scientific and educational resources*. Інформаційні технології і засоби навчання. 2010; 4(18): 35-39.
50. Лобузіна К. Технології організації знаннєвих ресурсів у бібліотечно-інформаційній діяльності. Київ: Національна академія наук України; 2012. 251 с.
51. Литвинова С. Особливості розробки критеріїв оцінювання електронних освітніх ресурсів. Наукові записки. Серія: проблеми методики фізико-математичної і технологічної освіти. 2013; 1(4): 63-67.

52. Марущак АІ. Інформаційні ресурси держави: зміст та проблема захисту. Правова інформатика. 2009; 1(21): 65-71.
53. Поморова ОВ, Говорущенко ТО. Проектування інтерфейсів користувача. Хмельницький: Хмельницький національний університет; 2011. 206 с.
54. 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. ІПРІ; 2017, с. 180-191.
55. Web Of Science [Internet], [cited 2018 June 16]. Available from: https://www.webofknowledge.com/
56. Scopus Preview [Internet], [cited 2018 June 16]. Available from: https://www.scopus.com/home.uri
57. Science Direct [Internet], [cited 2018 June 16]. Available from: https://www.sciencedirect.com/
58. 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. ІПРІ, 2016, с. 110-111.
59. Google Scolar [Internet], [cited 2018 June 16]. Available from: https://scholar.google.com/
60. Google Scholar – Top publications [Internet], [cited 2018 June 16]. Available from: https://scholar.google.com.ua/citations?view_op=top_venues&hl=en
61. Google Scholar - Antony van Raan profile [Internet], [cited 2018 June 16]. Available from: https://scholar.google.com.ua/citations?user=G-2AKcgAAAAJ&hl=en
62. Sun Y, Barber R, Gupta M, Aggarwal C, Han J. Co-author Relationship Prediction in Heterogeneous Bibliographic Networks. In: 2011 International Conference on Advances in Social Networks Analysis and Mining [Internet]; 2011 July 25-27;

Kaohsiung, Taiwan; 2011 Aug 18 [cited 2017 Jan 25]; Available from: https://ieeexplore.ieee.org/abstract/document/5992571/

63. Glänzel W, Schubert A. Analysing Scientific Networks Through Co-Authorship. In: Moed H.F., Glänzel W., Schmoch U. (eds) Handbook of Quantitative Science and Technology Research. Springer, Dordrecht. 2004: 257-276.
64. Haiyan H, Hildrun K, Zeyuan L. The structure of scientific collaboration networks in Scientometrics. Scientometrics. 2007. 75(2).
65. Google Scholar – Text Mining search results [Internet], [cited 2018 June 16]. Available from: https://scholar.google.com.ua/scholar?hl=en&as_sdt=0%2C5&q=text+mining&btnG=&oq=text
66. CSIRO [Internet], [cited 2018 June 16]. Available from: https://www.csiro.au/
67. Altmetric [Internet], [cited 2018 June 16]. Available from: https://www.altmetric.com/
68. European Commission – Joint Research Centre Publication Repository [Internet], [cited 2018 June 16]. Available from: https://ec.europa.eu/info/departments/joint-research-centre_en
69. The David and Yolanda Kats Faculty of Arts, Tel Aviv University [Internet], [cited 2018 June 16]. Available from: https://en-arts.tau.ac.il/
70. Cornel University – arviv.org [Internet], [cited 2018 June 16]. Available from: https://arxiv.org/
71. Zenodo [Internet], [cited 2018 June 16]. Available from: https://zenodo.org/
72. Springer [Internet], [cited 2018 June 23]. Available from: https://www.springer.com/gp
73. Elsevier [Internet], [cited 2018 June 23]. Available from: https://www.elsevier.com/
74. Inetrantional Journal of Scientific Engineering and Research [Internet], [cited 2018 June 23]. Available from: http://www.ijser.in/?gclid=Cj0KCQiA37HhBRC8ARIsAPWoO0wFOxIkLpLh8tEz-HLLHr3rAmyp3OUI9b_czb5XpyIc4ppguiTz0UaAskVEALw_wcB

75. Scientific Research – An Academic Publisher [Internet], [cited 2018 June 23]. Available from: https://www.scirp.org/
76. International Journal of Computer (IJC) [Internet], [cited 2018 June 23]. Available from: http://ijcjournal.org/index.php/InternationalJournalOfComputer/index
77. Open Science Journal [Internet], [cited 2018 June 23]. Available from: https://osjournal.org/submissions.html?gclid=Cj0KCQiApvbhBRDXARIsALnNoK271H7iB1rUYKLOPlkjRfbhmIOdsd3OH766u-TK8Dy2m3Oda0YsIMwaAjiKEALw_wcB
78. Open AIRE [Internet], [cited 2018 June 23]. Available from: https://www.openaire.eu/
79. Sofware Heritage [Internet], [cited 2018 June 23]. Available from: https://www.softwareheritage.org/
80. GitHub [Internet], [cited 2018 June 23]. Available from: https://github.com
81. Debain – The Universal Operating System [Internet], [cited 2018 June 23]. Available from: https://www.debian.org/
82. Gitorious Valhalla [Internet], [cited 2018 June 23]. Available from: https://gitorious.org
83. Google Code [Internet], [cited 2018 June 23]. Available from: code.google.com/
84. GNU Operating System [Internet], [cited 2018 June 23]. Available from: https://www.gnu.org/home.en.html
85. CERN Scientific Information Service [Internet], [cited 2018 June 23]. Available from: http://library.cern/
86. Registry of Research Data Repositories [Internet], [cited 2018 June 23]. Available from: https://www.re3data.org/
87. Андрущенко ВБ. Підходи до визначення критеріїв для аналізу on-line ресурсів наукової інформації. Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія «Технчні науки». 2018. 29(68), 4:84-89.

88. Боярский КК. Введение в компьютерную лингвистику. СПбЖ НИУ ИТМО; 2013. 72с.

89. Ланде ДВ, Андрущенко ВБ. Побудова мережі предметних областей на базі ресурсу arxiv. Реєстрація, зберігання і обробка даних. 2018. 20(2): 12-22.

90. Андрущенко ВБ, Ланде ДВ. Побудова онтології за допомогою сканування ресурсів Wikipedia. В: Литвиненко ОЄ. Матеріали VIII Міжнародної науково-технічної конференції «Інтелектуальні технології лінгвістичного аналізу»; 25-26 жовтня 2016 року; Національний авіаційний університет. Київ. Київ: НАУ; 2016, с. 10.

91. 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.

92. Lande DV, AndrushchenkoVB, Balagure IV. Formation of the Subject Area on the Base of Wikipedia Service. В: Голенков ВВ. Материалы Міжнародної конференції Open Semantic Technologies for Intelligent Systems; 17-19 лютого 2017 року. Білоруський державний інститут інформатики та радіоелектроніки; БДУІР;2017, с. 211-215.

93. Штефан СВ. Статистичні методи досліджень. Інститут журналістики [Інтернет]. Київ: Інститут журналістики. [цитовано 23 травня 2018]ю Доступно: http://journlib.univ.kiev.ua/navch/StatMetodyDoslid.pdf

94. Гузь АН. Анализ систем оценок научных публикаций. Киев: Институт механики им. С.П. Тимошенко НАНУ; 2013. 280 с.

95. Акоев МА, Маркусова ВА, Москалева ОВ, Писляков ВВ. Руководство по наукометрии: индикаторы развития науки и технологии. Екатеринбург: ИПЦ УрФУ; 2014. 250 с.

96. Снарский АА, Ландэ ДВ. Моделирование сложных сетей. Киев: Инжиниринг; 2015. 212 с.

97. Ланде ДВ. Створення термінологічної моделі предметної області шляхом зондування Google Scholar Citations. Правова інформатика. 2015; 2(46):3-8.
98. Ланде ДВ, Елементи комп'ютерної лінгвістики в правовій. Київ: НДІІП НАПрН України; 2014. 168 с.
99. 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.
100. 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.
101. 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
102. Ландэ ДВ, Снарский АА. Подход к созданию терминологических онтологий. Онтология проектирования. 2014; 2(12): 83-91.
103. Ланде ДВ, Андрущенко ВБ, Балагура ІВ. Вікі-індекс популярності авторів наукових публікацій. Реєстрація, зберігання і обробка даних. 2016. 18 (4): с. 44-54.
104. 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.
105. Lande DV, Andrushchenko VB, Balagura IV. Wiki-index of authors popularity. Arxiv.org. 1702.04614. 2017.Available from: https://arxiv.org/abs/1702.04614

106. Brezina V. Use of Google Scholar in corpus-driven EAP researchю. Journal of English for Academic Purposes. 2012; 11: 319-331.

107. Добров БВ, Соловьев ВД, Лукашевич НВ, Иванов ВВ. Онтологии и тезаурусы. Модели, инструменты, приложения. Москва: Бином; 2009. 173 с.

108. Ландэ ДВ, Снарский АА, Безсуднов ИВ. Интернетика : навигация в сложных сетях : модели и алгоритмы. Москва: Либроком; 2009. 264 с.

109. Sorokina D, Gehrke J, Warner S. Plagiarism Detection in arXiv. In: Sixth International Conference on Data Mining (ICDM'06) [Internet]; 2009 Dec 18-22; Hong Cong, China. IEEE; 2007 Feb 8 [cited 2018 Apr 12]. Available from: https://ieeexplore.ieee.org/document/4053155/

110. Warner S. Open Archives Initiative protocol development and implementation at arXiv. Arxiv.org. arXiv:cs/0101027. 2001. Available from: https://arxiv.org/abs/cs/0101027

111. Юрченко ОВ. Дефініція концепту в сучасних лінгвістичних дослідженнях. Вісник Запорізького національного університету. 2008; 2: 268-271.

112. Андрущенко ВБ. Нові інформаційні технології пошуку і обробки даних ресурсу препринтів arxiv. Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія «Технчні науки». 2018. 29(68), 3:84-89.

113. Common Evaluation Measures. Text Retrieval Conference [Internet]. [cited 2018 March]. Available from: https://trec.nist.gov/pubs/trec16/appendices/measures.pdf

114. Агеев М, Кураленок И, Некрестьянов И. Официальніе метрики РОМИП 2010. В: Труды РОМИП-2010 [Интернет]; 2010 Окт 15; Казань [цитовано 2018 Бер 23]; с. 172-187. Доступно: http://romip.ru/romip2010/20_appendix_a_metrics.pdf

115. 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.

116. Ланде ДВ, Андрущенко ВБ, Wikipedia Index of Scientist’s Popularity. В: Дичка ІА. XVII Міжнародна наукова конференція імені Т.А. Таран "Інтелектуальний аналіз інформації. ІАІ2017, Київ, 17-19 травня 2017 р. Просвіта, 2017. с. 137-143.

117. 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.

118. Lande D, AndrushchenkoV, 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.

ДОДАТОК А Результати аналізу систем наукової інформації з урахуванням запропонованих критеріїв

Таблиця А.1 – Узагальнені результати аналізу систем наукової інофрмації з урахуванням запропонованих критеріїв

№ п/п	Назва ресурсу, посилання	К	Відкритий доступ / Передплата	Необхідність реєстрації на ресурсі для доступу до інформації	Предметна область	Науковий напрямок	Назва публікації/ експерименту/ проекту	Автор/автори	Ключові слова	Реферат/ анотація	Повний текст публікації/опис наукових результатів, експерименту	Посилання на повний текст/опис наукових результатів, експерименту	Наукометричні показники	Пошук інформації на головній сторінці	Можливість збереження представленої інформації
	S		K1	K2	K3	K4	K5	K 6	K7	K8	K9	K10	K11	K12	K13
1.	Web of Science www.webofknowledge.com/	S1	0	1	1	1	1	1	1	1	0	0	1	1	1
2.	Scopus www.scopus.com	S2	0	1	1	1	1	1	1	1	0	0	1	1	1
3.	Google Scholar https://scholar.google.com.ua/	S3	1	1	0	1	1	1	0	1	0	1	1	1	0
4.	Research Publications Repository https://publications.csiro.au	S4	1	1	0	1	1	1	1	1	0	1	0	1	0
5.	Joint Research Centre Publications Repository http://publications.jrc.ec.europa.eu/repository/	S5	1	1	0	0	1	1	0	1	0	1	0	1	0
6.	Research and Publications Archive https://en-arts.tau.ac.il	S6	1	1	1	1	0	0	0	0	1	1	0	1	0

Продовження таблиці А.1

№ п/п	Назва ресурсу, посилання	К	Відкритий доступ / Передплата	Необхідність реєстрації на ресурсі для доступу до інформації	Предметна область	Науковий напрямок	Назва публікації/ експерименту/ проекту	Автор/ автори	Ключові слова	Реферат/ анотація	Повний текст публікації/опис наукових результатів, експерименту	Посилання на повний текст/опис наукових результатів, експерименту	Наукометричні показники	Пошук інформації на головній сторінці	Можливість збереження представленої інформації
7.	ArXiv https://arxiv.org/	S7	1	1	1	1	1	1	0	1	0	1	0	1	0
8.	Zenodo https://zenodo.org/	S8	1	1	0	1	1	1	1	1	1	0	0	1	0
9.	Springer www.springer.com	S9	1	1	0	1	1	1	0	1	0	1	0	1	0
10.	Elsevier www.elsevier.com	S10	1	1	0	0	1	0	0	1	0	0	1	0	0
11.	International Organization of Scientific Research http://iosrjournals.org/	S11	1	1	0	1	0	0	0	0	0	0	1	0	0
12.	Scientific Research An academic Publisher http://www.scirp.org	S12	1	1	1	1	1	1	0	1	0	1	0	1	0
13.	International Journal of Computer http://ijcjournal.org	S13	1	1	0	1	1	1	1	1	1	1	0	1	0
14.	Open Science Journal https://osjournal.org	S14	1	1	0	0	1	1	1	1	0	1	0	1	0

Кінець таблиці А.1

№ п/п	Назва ресурсу, посилання	К	Відкритий доступ / Передплата	Необхідність реєстрації на ресурсі для доступу до інформації	Предметна область	Науковий напрямок	Назва публікації/ експерименту/ проекту	Автор/автори	Ключові слова	Реферат/ анотація	Повний текст публікації/опис наукових результатів, експерименту	Посилання на повний текст/опис наукових результатів, експерименту	Наукометричні показники	Пошук інформації на головній сторінці	Можливість збереження представленої інформації
15.	OpenAire www.openaire.eu	S15	1	1	0	1	1	1	0	0	0	1	0	0	0
16.	Software Heritage https://www.softwareheritage.org/	S16	1	1	0	0	0	0	0	0	0	1	0	0	0
17.	CERN Scientific Infromation Service http://library.cern	S17	1	1	0	1	1	1	1	1	0	1	0	1	0
18.	Registry of Research Data Repositories www.re3data.org	S18	1	1	1	1	1	0	1	1	0	1	0	1	0

Додаток Б. Лістінги програми реалізації алгоритму побудови мовою Java онтології на базі он-лайн енциклопедії Вікіпедія

```
package graph;


public class Edge {
	private int mSourceId;
	private int mTargetId;


	public Edge(int source, int target) {
		mSourceId = source;
		mTargetId = target;
		System.out.println("Add edge " + toString());
	}


	public int getSource() {
		return mSourceId;
	}


	public int getTarget() {
		return mTargetId;
	}


	public String toString() {
		return "source = " + mSourceId + " target = " + mTargetId;
	}
}


package graph;
import java.util.ArrayList;
```

```
import java.util.List;

public class GraphVertex {
	/* wiki base address, e.g. https://en.wikipedia.org */
	private String mWikiBaseURL;
	/* current wiki article wrt wiki base address, e.g. wiki/Science */
	private String mCurrentWikiArticle;
	/* base notion which must be found within the current article */
	private String mBaseNotion;
	private Storage mStorage;

	public GraphVertex(String wikiBase, String baseNotion, String current) {
		mWikiBaseURL = wikiBase;
		mBaseNotion = baseNotion;
		mCurrentWikiArticle = current;
		mStorage = new Storage();
	}

	public void run() {
		System.out.println("Using " + mWikiBaseURL + "/" +
					mCurrentWikiArticle + " as root URL, base notion is " +
					mBaseNotion);
		Node initialNode = mStorage.addNode(null, mBaseNotion,
mCurrentWikiArticle);
		WikiPage page = new WikiPage(mWikiBaseURL, mBaseNotion,
mCurrentWikiArticle, initialNode);
		page.run(mStorage);
		/* get file name from the wiki article path, b/c base notion may
		 * not uniquely identify the graph */
		String fName = mCurrentWikiArticle.replace('/', '_');
		if (fName.startsWith("_")) {
			fName = fName.substring(1);
		}
		mStorage.csvExport(fName);
	}
}

package graph;

public class Main {
	private static String WIKI_DEFAULT_URL = "https://en.wikipedia.org";
	private static String GRAF_VERTEX_DEFAULT = "/wiki/Scientometrics";
	private static String NOTION_DEFAULT = "Scientometrics";

	public static void main(String[] args) {
		String wiki_url = WIKI_DEFAULT_URL;
		String graf_vertex = GRAF_VERTEX_DEFAULT;
		String notion = NOTION_DEFAULT;

		if (args.length > 1) {
			graf_vertex = args[0];
			if (!graf_vertex.startsWith("/wiki")) {
				graf_vertex = "/wiki/" + graf_vertex;
```

```
                        }
                        notion = args[1];
                }
                if (args.length > 2) {
                        wiki_url = args[2];
                }
                new GraphVertex(wiki_url, notion, graf_vertex).run();
        }
}


package graph;

public class Node {
        private static int mNextId = 0;
        private String mTitle;
        private String mURL;
        private int mId;

        public Node(String title, String href) {
                mId = mNextId++;
                mTitle = title;
                mURL = href;
                System.out.println("Add node " + toString());
        }

        public String getURL() {
                return mURL;
        }

        public int getId() {
                return mId;
        }

        public String getTitle() {
                return mTitle;
        }

        public String toString() {
                return "id = " + mId + " \"" +  mTitle + "\" " + mURL;
        }
}
```

```
package graph;

import java.io.FileNotFoundException;
import java.io.IOException;
import java.io.PrintWriter;
import java.util.ArrayList;
import java.util.List;

public class Storage {
	private List<Node> mNodes;
	private List<Edge> mEdges;

	public Storage() {
		mNodes = new ArrayList<Node>();
		mEdges = new ArrayList<Edge>();
	}

	/*
	 * Look up or add a node:
	 *  - return true if node has not been found and added to the list
	 *    this means that we must continue parsing pages
	 *  - return false if node exists, just add edge
	 */
	public Node addNode(Node sourceNode, String title, String href) {
		for (Node i : mNodes) {
			if (i.getURL().equals(href)) {
				/* found in the storage, add edge */
				System.out.println("Node found in the storage: " + href);
				mEdges.add(new Edge(sourceNode.getId(), i.getId()));
				return null;
			}
		}
		/* node not found, add new */
		System.out.println("Node NOT found in the storage: " + href);
		Node newNode = new Node(title, href);
		mNodes.add(newNode);
		if (sourceNode != null) {
			mEdges.add(new Edge(sourceNode.getId(), newNode.getId()));
		}
		return newNode;
	}

	public void csvExport(String baseName) {
		PrintWriter out;
		try {
			out = new PrintWriter(baseName + "_nodes.csv");
			out.println("Id,Label");
			for (Node i : mNodes) {
				out.println(i.getId() + "," + i.getTitle());
			}
			out.close();
		} catch (IOException e) {
			e.printStackTrace();
```

```
                }
                try {
                        out = new PrintWriter(baseName + "_edges.csv");
                        out.println("Source,Target");
                        for (Edge i : mEdges) {
                                out.println(i.getSource() + "," + i.getTarget());
                        }
                        out.close();
                } catch (IOException e) {
                        e.printStackTrace();
                }
        }
}


package graph;

import java.io.IOException;
import java.util.ArrayList;
import java.util.List;
import java.util.regex.Matcher;
import java.util.regex.Pattern;

import org.jsoup.Jsoup;
import org.jsoup.nodes.Document;
import org.jsoup.nodes.Element;
import org.jsoup.select.Elements;

public class WikiPage {
        /* wiki base address, e.g. https://en.wikipedia.org */
        private String mWikiBaseURL;
        /* current wiki article wrt wiki base address, e.g. wiki/Science */
        private String mCurrentWikiArticle;
        /* base notion which must be found within the current article */
        private String mBaseNotion;
        /* this wiki page node */
        private Node mSourceNode;

        public WikiPage(String wikiBaseURL, String baseNotion, String currentArticle,
                        Node sourceNode) {
                mWikiBaseURL = wikiBaseURL;
                mBaseNotion = baseNotion;
                mCurrentWikiArticle = currentArticle;
                mSourceNode = sourceNode;
        }

        public void run(Storage storage) {
                System.out.println("Iterating over " + mWikiBaseURL + mCurrentWikiArticle);
                try {
                        Document doc = Jsoup.connect(mWikiBaseURL +
mCurrentWikiArticle).get();

                        Elements elements = doc.body().getElementsByTag("p");
```

```
                        /* get text of all paragraphs into a single string */
                        String text = "";
                        for (Element p : elements) {
                                text += p.text();
                        }
                        /* check if page contains base notion (base notion can
                         * be a multi-word text) */
                        String patternString = "(\\s|)" + mBaseNotion + "(\\s|)";
                        Pattern pattern = Pattern.compile(patternString,
Pattern.CASE_INSENSITIVE);
                        Matcher matcher = pattern.matcher(text);
                        if (!matcher.find()) {
                                System.out.println(mBaseNotion + " not found in this page");
                                return;
                        }
                        for (Element p : elements) {
                                Elements hrefs = p.select("a");
                                for (Element a : hrefs) {
                                        if (a.attr("href").startsWith("#cite_note")) {
                                                /* skip references */
                                                continue;
                                        }
                                        String currentTitle = a.attr("title");
                                        String currentArticle = a.attr("href");
                                        if (!currentArticle.startsWith("/wiki")) {
                                                /* skip non wiki links */
                                                continue;
                                        }
                                        if (currentArticle.contains("Special:BookSources"))
{
                                                /* skip book ISDN links */
                                                continue;
                                        }
                                        System.out.println("\nNow at " + mWikiBaseURL
+ mCurrentWikiArticle +
                                                        " node id = " +
mSourceNode.getId());
                                        Node newNode = storage.addNode(mSourceNode,
currentTitle, currentArticle);
                                        if (newNode != null) {
                                                WikiPage page = new
WikiPage(mWikiBaseURL, mBaseNotion,
                                                                currentArticle,
newNode);
                                                page.run(storage);
                                        }
                                }
                        }
                } catch (IOException e) {
                        System.out.print("Failed to retrieve " + mWikiBaseURL + "/" +
mCurrentWikiArticle +
                                        ": " + e.getMessage());
                        e.printStackTrace(); } }}
```

ДОДАТОК В Словники предметних областей для ресурсу препринтів arXiv, які містять переліки наукових напрямків, що конкретизують відповідну предметну область.

1. Computer Science

- Computing Research Repository
- Artificial Intelligence
- Computation and Language
- Computational Complexity
- Computational Engineering
- Finance, and Science
- Computational Geometry
- Computer Science and Game Theory
- Computer Vision and Pattern Recognition
- Computers and Society
- Cryptography and Security
- Data Structures and Algorithms
- Databases
- Digital Libraries
- Discrete Mathematics
- Distributed, Parallel, and Cluster Computing
- Emerging Technologies
- Formal Languages and Automata Theory
- General Literature
- Graphics
- Hardware Architecture
- Human-Computer Interaction
- Information Retrieval
- Information Theory
- Learning
- Logic in Computer Science
- Mathematical Software
- Multiagent Systems
- Multimedia
- Networking and Internet Architecture
- Neural and Evolutionary Computing
- Numerical Analysis
- Operating Systems
- Other Computer Science
- Performance
- Programming Languages
- Robotics
- Social and Information Networks
- Software Engineering
- Sound

- Symbolic Computation
- Systems and Control

2. Economics

- Econometrics

3. Electrical Engineering and System Science

- Audio and Speech Processing
- Image and Video Processing
- Signal Processing

4. Mathematics

- Algebraic Geometry
- Algebraic Topology
- Analysis of PDEs
- Category Theory
- Classical Analysis and ODEs
- Combinatorics
- Commutative Algebra
- Complex Variables
- Differential Geometry
- Dynamical Systems
- Functional Analysis
- General Mathematics
- General Topology
- Geometric Topology
- Group Theory
- History and Overview
- Information Theory
- K-Theory and Homology
- Logic
- Mathematical Physics
- Metric Geometry
- Number Theory
- Numerical Analysis
- Operator Algebras
- Optimization and Control
- Probability
- Quantum Algebra
- Representation Theory
- Rings and Algebras
- Spectral Theory

- Statistics Theory
- Symplectic Geometry

5. Physics

- Astrophysics
- Astrophysics of Galaxies
- Cosmology and Nongalactic Astrophysics
- Earth and Planetary Astrophysics
- High Energy Astrophysical Phenomena Instrumentation and Methods for Astrophysics
- Solar and Stellar Astrophysics
- Condensed Matter
- Disordered Systems and Neural Networks
- Materials Science
- Mesoscale and Nanoscale Physics
- Other Condensed Matter
- Quantum Gases
- Soft Condensed Matter
- Statistical Mechanics
- Strongly Correlated Electrons
- Superconductivity
- General Relativity and Quantum Cosmology
- High Energy Physics - Experiment
- High Energy Physics - Lattice
- High Energy Physics - Phenomenology
- High Energy Physics - Theory
- Mathematical Physics
- Nonlinear Sciences
- Adaptation and Self-Organizing Systems
- Cellular Automata and Lattice Gases
- Chaotic Dynamics
- Exactly Solvable and Integrable Systems
- Pattern Formation and Solitons
- Nuclear Experiment
- Nuclear Theory
- Physics
- Accelerator Physics
- Applied Physics
- Atmospheric and Oceanic Physics
- Atomic Physics
- Atomic and Molecular Clusters
- Biological Physics
- Chemical Physics
- Data Analysis
- Statistics and Probability

- Fluid Dynamics
- General Physics
- Geophysics
- History and Philosophy of Physics
- Instrumentation and Detectors
- Medical Physics
- Optics
- Physics Education
- Physics and Society
- Plasma Physics
- Popular Physics
- Space Physics
- Quantum Physics

6. Quantitative biology

- Biomolecules
- Cell Behavior
- Genomics
- Molecular Networks
- Neurons and Cognition
- Other Quantitative Biology
- Populations and Evolution
- Quantitative Methods
- Subcellular Processes
- Tissues and Organs

7. Quantitative Finance

- Computational Finance
- Economics
- General Finance
- Mathematical Finance
- Portfolio Management
- Pricing and Securities
- Risk Management
- Staticstical Finance
- Trading and Market Microsrtructure

8. Statistics

- Machine learning
- Methodology
- Other Statistics

ДОДАТОК Г Список публікацій здобувача

1. Ланде ДВ, Андрущенко ВБ, Балагура ІВ. Вікі-індекс популярності авторів наукових публікацій. Реєстрація, зберігання і обробка даних. 2016. 18 (4): с. 44-54. (Особистий внесок – оцінка можливостей використання бібліографічних посилань статей Wikipedia для виокремлення інформації про авторів).

2. Ланде ДВ, Андрущенко ВБ. Побудова мережі предметних областей на базі ресурсу arxiv. Реєстрація, зберігання і обробка даних. 2018. 20(2): 12-22. (Особистий внесок – розробка алгоритму пошуку наукових публікацій за заданим концептом і формування мережі предметних областей на базі отриманих результатів, оцінка отриманих результатів)
3. Андрущенко ВБ. Нові інформаційні технології пошуку і обробки даних ресурсу препринтів arxiv. Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія «Технчні науки». 2018. 29(68), 3:84-89. (Особистий внесок – розробка моделі «Концепт – система наукових напрямків, розробка алгоритму для реалізації побудови мережі предметних областей, апробація результатів на заданому концепті)
4. Андрущенко ВБ. Підходи до визначення критеріїв для аналізу on-line ресурсів наукової інформації. Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія «Технчні науки». 2018. 29(68), 4:88-95. (Особистий внесок – розробка підходів до формування критеріїв для оцінки наукової інформації)
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. (Особистий внесок – визначення підходів до оцінки активності вітчизняних вчених за рахунок сканування великих масивів інформації про грантові пропозиції)
6. Андрущенко ВБ, Кияк БР. Обґрунтування критеріїв оцінювання фундаментальних наукових досліджень. Наука та наукознавство. Грудень

2015; 4(89):67-72. (Особистий внесок – розробка критеріїв із запропонованого переліку щодо оцінювання фундаментальних досліджень)

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. (Особистий внесок – участь у розробці алгоритму реалізації поставленої задачі)
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 фундамнтальних досліджень (досвід Державного фонду фундаментальних досілджень України); с. 74-81. (Особистий внесок – моніторинг масиву даних про проекти спільних конкурсів та виокремлення наукометричної інформації із провденням аналізу отриманих результатів)
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. (Особистий внесок – аналіз масиву даних про конкурсні проекти за заданим напрямком, викоремлення наукометричних даних результатів пошуку та їх аналіз)
10. Андрущенко ВБ. Інформаційно-аналітична діяльність Державного фонду фундментальних 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. (Особистий внесок – опис процесу розробки, вдосконалення та порядку роботи інформаційно-аналітичної системи ДФФД, підходи до аналізу даних, що містить система)
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. (Особистий внесок – участь у розробці алгоритму для реалізації поставленої задачі)

12. Андрущенко ВБ, Ланде ДВ. Побудова онтології за допомогою сканування ресурсів Wikipedia. В: Литвиненко ОЄ. Матеріали VIII Міжнародної науково-технічної конференції «Інтелектуальні технології лінгвістичного аналізу»; 25-26 жовтня 2016 року; Національний авіаційний університет. Київ. Київ: НАУ; 2016, с. 10. (Особистий внесок – розробка та програмна реалізація алгоритму для вирішення задачі)
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. (Особистий внесок – виокремлення та аналіз результатів за заданою тематикою розробленими програмними засобами)
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. ІПРІ, 2016, с. 110-111. (Особистий внесок – розробка підходів до порівнння інформаційних систем наукометричної іфнормації)
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. (Особистий внесок – оцінка масиву інформації, що містить система та розробка алгоритмів для виокремлення інформації для формування нових масивів даних)

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. (Особистий внесок – оцінка масиву інформації, що містить система та розробка алгоритмів для виокремлення інформації для формування нових масивів даних)
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. ІПРІ; 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.

(Особистий внесок – визначення параметрів для проведення моніторингу, аналізу та узагальнення інформації про ресурси мережі Інтернет, що містять наукову інформацію)
18. Lande DV, AndrushchenkoVB, Balagure IV. Formation of the Subject Area on the Base of Wikipedia Service. В: Голенков ВВ. Материалы Міжнародної конференції Open Semantic Technologies for Intelligent Systems; 17-19 лютого 2017 року. Білоруський державний інститут інформатики та радіоелектроніки; БДУІР;2017, с. 211-215. (Особистий внесок – розробка алгоритму та його програмна реалізація)
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. (Особистий внесок – участь у розробці алгоритму для реалізації поставленої задачі)

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. (Особистий внесок – розробка моделі, алгоритму, програмна реалізація та апробація результатів дослідження)
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. ІПРІ; 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.

    (Особистий внесок – моніторинг системи наукометричної інформації для виокремлення масиву публікацій за заданим напрямком та подальший аналіз розробленими програмними засобами)
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. (Особистий внесок – участь у розробці алгоритму для реалізації поставленої задачі)
23. Lande D, AndrushchenkoV, 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. (Особистий внесок – розробка алгоритму та апробація результатів дослідження)

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 (Особистий внесок – участь у розробці алгоритму для реалізації поставленої задачі)
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. (Особистий внесок – участь у розробці алгоритму для реалізації поставленої задачі)

ДОДАТОК Д Апробація результатів

1. Міжнародна науково-технійчна конференція «GRANT-2015», 1-2 жовня 2015 року, Мінстерство освіти і науки України, Київ, Україна, доповідь.
2. Міжнарожна науково-практивна конференція «Відкриті семантичні технології для інлелектуальних систем» OSTIS-2016, Білоруський державний університет інформатики та радіоелектроніки, Республіка Білорусь, Мінськ, спільна доповідь.
3. Міжнарожна науково-практивна конференція «Відкриті семантичні технології для інлелектуальних систем» OSTIS- 2017, Білоруський державний університет інформатики та радіоелектроніки, Республіка Білорусь, Мінськ, спільна доповідь.
4. 2-а Міжнародна науково-практична конференція «Наукова комунікація в цифрову епоху», 30 січня – 1 лютого 2014 року, Національний університет «Києво-Могилянська академія», Київ, Україна, доповідь.
5. 3-я Міжнародна науково-практична конференція «Наукова комунікація в цифрову епоху», 10-12 березня 2015 року, Національний університет «Києво-Могилянська академія», Київ, Україна, доповідь.
6. 4-а Міжнародна науково-практична конференція «Наукова комунікація в цифрову епоху», 30-31 березня 2016 року, Національний університет «Києво-Могилянська академія», Київ, Україна, доповідь.
7. 6-а Міжнародна науково-практична конференція «Наукова комунікація в цифрову епоху», 29-30 березня 2018 року, Національний університет «Києво-Могилянська академія», Київ, Україна, доповідь.
8. 18-та Міжнародна конференція «Системний аналіз та інформаційні технології» (SAIT-2016), 30 травня-2 червня 2016 року, Інститут прикладного системного аналізу НТУУ «КПІ імені Ігоря Сікорського», Київ, Україна, доповідь.
9. 8-а Науково-практична конференція «Наука України у світовому інформаційному просторі», травнь 2016 року, Київ, Україна, доповідь.

10. Щорічна підсумкова конференція «Реєстрація, зберігання та обробка даних», 15-16 травня 2016, Інститут проблем реєстрації інформації НАН України, Київ, Україна, постерна доповідь.
11. Щорічна підсумкова конференція «Реєстрація, зберігання та обробка даних», 17-18 травня 2017, Інститут проблем реєстрації інформації НАН України, Київ, Україна, доповідь.
12. IX Міжнародна науково-технічна конференція «Інтелектуальні технології лінгвістичного аналізу», 25-26 жовтня 2016 року, Національний авіаційний університет, Київ, Україна, доповідь.
13. X Міжнародна науково-технічних конференціях «Інтелектуальні технології лінгвістичного аналізу», 24-25 жовтня 2017 року, Національний авіаційни університет, Київ, Україна, доповідь.
14. XVI Міжнародна науково-практична конференція «Інформаційні технології та безпека ІТБ-2016», 1 грудня 2016 року, Інститут проблем реєстрації інформації НАН України, Київ, Україна, доповідь.
15. XVI Міжнародна науково-практична конференція «Інформаційні технології та безпека ІТБ-2016», 1 грудня 2016 року, Інститут проблем реєстрації інформації НАН України, Київ, Україна, доповідь.
16. XVII Міжнародна науково-практична конференція «Інформаційні технології та безпека ІТБ-2017», 30 листопада 2017 року, Інститут проблем реєстрації інформації НАН України, Київ, Україна, доповідь.
17. Міжнародна конференції «IEEE Second International Conference on Data Stream Mining and Processing», 21-25 серпня 2018 року, Львів, Україна, доповідь.

ДОДАТОК Е

Довідки про впровадження результатів дисертаційної роботи

Підприємство Української академії наук
«Інститут системних досліджень та інформаційних технологій»

Україна, 03142 м. Київ, вул. Семашка, 13

Довідка
Про впровадження результатів дисертаційної роботи
Андрущенко Валентини Борисівни
«Інформаційні технології наукометричного аналізу на основі моніторингу ресурсів мережі Інтернет»
поданої на здобуття наукового ступеню кандидата технічних наук за спеціальністю 05.13.06 – Інформаційні технології

Результати дисертаційної роботи Андрущенко Валентини Борисівни «Інформаційні технології наукометричного аналізу на основі моніторингу ресурсів мережі Інтернет» було використано при формуванні стратегії підготовки грантового дослідження в рамках різноманітних конкурсних програм, зокрема – пошуку партнерів для формування наукових колаборацій. Поряд із підходами, які розроблені та вже використовуються Інститутом (або організацією), запропоновані в дисертаційній роботі програмні додатки дозволяють розширити низку способів реалізації залучення партнерів за рахунок пошуку відповідної інформації з наукової точки зору із використанням бібліометричної інформації різних он-лайн джерел для подальшого аналізу та формування проектних пропозицій.

Довідка надана для подання у Спеціалізовану вчену раду К58.052.06.

Директор УАН «ІСДІТ» /Посдинок Н.Л./

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ «КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ СІКОРСЬКОГО»

ФАКУЛЬТЕТ ПРИКЛАДНОЇ МАТЕМАТИКИ

КАФЕДРА ПРИКЛАДНОЇ МАТЕМАТИКИ

03056, м. Київ, пр-т Перемоги, 37; тел. (+38 044) 236-95-05
http://pma.fpm.kpi.ua ел. адреса: pmakpi@gmail.com

14.01.2019 № 5/1910
на № _____

Довідка
про впровадження результатів дисертаційної роботи
Андрущенко Валентини Борисівни
«Інформаційні технології наукометричного аналізу на основі моніторингу ресурсів мережі інтернет»,
поданої на здобуття наукового ступеню кандидата технічних наук за спеціальністю 05.13.06 – Інформаційні технології

Результати дисертаційної роботи Андрущенко Валентини Борисівни «Інформаційні технології наукометричного аналізу на основі моніторингу ресурсів мережі Інтернет» було використано при формуванні матеріалів щодо розвитку навичок студентів та викладачів кафедри прикладної математики НТУУ «КПІ імені Ігоря Сікорського» основним елементам підготовки грантових запитів у рамках програми підтримки досліджень та інновацій «Горизонт-2020», зокрема – пошуку партнерів для наукової співпраці та формування відповідних консорціумів. Крім традиційних підходів, таких як онлайн-ресурси для пошуку партнерів, запропоновані в дисертаційній роботі програмні застосунки дозволяють реалізувати пошук із використанням бібліометричної інформації різних онлайн-джерел для подальшого зіставлення й формування стратегії дослідження.

Довідка надана для подання у Спеціалізовану вчену раду К58.052.06.

Завідувач кафедри прикладної математики
факультету прикладної математики
НТУУ «Київський політехнічний інститут
імені Ігоря Сікорського»,
доктор технічних наук, професор О. Р. Чертов