УДК 004.656

Андрущенко В.Б.
Інститут проблем реєстрації інформації НАН України

НОВІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ПОШУКУ Й ОБРОБКИ ДАНИХ РЕСУРСУ ПРЕПРИНТІВ ARXIV

Стаття присвячена новим технологіям роботи з ресурсом відкритого доступу – архівом препринтів arXiv. За рахунок представлення даних, що містить ресурс у вигляді моделі «Концепт – система наукових напрямків», розроблено та представлено алгоритм роботи з ресурсом і виокремлено нові масиви інформації, які було сформовано шляхом застосування методів text mining до результатів пошуку на ресурсі за заданим концептом. У статті подано результати апробації алгоритму, їх візуалізація у середовищі Gephi та запропоновано підходи до інтерпретації отриманих даних. Також було представлено шляхи розвитку реалізованої задачі.

Ключові слова: *архів препринтів, концепт, науковий напрямок, публікація, граф.*

Постановка проблеми. Сьогодні наукометрія є одним з основних і допоміжних інструментів для організації та проведення аналізу результативності наукової діяльності окремого вченого, творчого колективу та організації. Вона є мірилом визначення успішності реалізації грантових програм як у межах країни, так і в світовому науковому просторі.

Згідно з Законом України «Про наукову та науково-технічну діяльність» (ст. 11, п. 2, пп. 3), оцінка наукової та науково-технічної діяльності й атестація здійснюються з урахуванням наукометричних показників.

Широкий спектр наукометричних показників дозволяє провести оцінку наукової роботи шляхом моніторингу публікаційної активності та цитованості. Загальноприйнятими у світі є два провідних наукометричних ресурси – «Scopus» (компанія – Elsevier, Нідерланди) та «Web Of Science» (компанія – Clarivate Analytics, Сполучені Штати Америки). Варто підкреслити, що доступ до цих ресурсів є передплатним. Натомість більшість наукових установ та освітніх закладів для оцінки наукового складника організації звертаються до показників наукометричного ресурсу компанії Google (Сполучені Штати Америки) – Google Scholar.

Водночас актуальним залишається отримання додаткової інформації та розробка нового інструментарію на базі наукометричних ресурсів і ресурсів наукової інформації відкритого доступу, не тільки для оцінки та моніторингу діяльності вчених, колективів та установ, а і, за рахунок синергії даних, для отримання нових підходів для пошуку міждисциплінарних зав'язків, що є одним із актуальних засобів формування міжнародних колаборацій, привернення уваги міжнародної наукової спільноти до наукових досліджень України та спільної реалізації грантових програм.

Також варто зауважити, що значна увага приділена розробці підходів для оцінки стану науки, аналізу діяльності науковців і методологічних підходів до розробки стратегій розвитку наукової сфери в Україні.

Одним із найбільш популярних ресурсів для розміщення результатів досліджень є ресурс препринтів Корнуельської бібліотеки – arXiv – найбільший архів електронних публікацій і їх препринтів відкритого доступу.

Репозитарій було запроваджено у 1991 р. Роботу ресурсу було спрямовано на розміщення публікацій, підготовлених до друку, за напрямком «Фізика», але сьогодні ресурс постійно розширюється, додаються нові розділи та відповідні підрозділи з інших наукових напрямків. arXiv є допоміжним інструментарієм для науковців в усьому світі. Ресурс є актуальним інструментом для користувачів із країн з обмеженим доступом до наукової інформації за рахунок можливості користування дзеркалами.

Аналіз останніх досліджень і публікацій. Досі всю увагу до ресурсу та досліджень на базі нього було зосереджено на впровадженні протоколів відкритого доступу на базі архіву препринтів [1] і способам виявлення плагіату в рамках ресурсу [2].

Водночас arXiv становить базу наукових препринтів та опублікованих праць за різними науковими напрямками, що оновлюється щоденно і містить найновіші результати досліджень із різних галузей знань.

Ресурсом передбачено процедура схвалення (endorsement) статті перед опублікуванням із залучення експертів із 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.

Постановка завдання. 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; відтворення у вигляді схематичних зображень результатів пошуку і проведення їх оцінки та подальшої інтерпретації.

Під концептом розуміємо [3] те, що називає зміст поняття, у цьому разі – зображення і зміст наукового явища, його характеристику, ознаку. Заданий для пошуку концепт може бути не тільки словом або словосполученням, що визначають науковий термін, характеристику наукового процесу тощо, а й власними іменами та назвами.

Сьогодні автором було запропоновано та реалізовано задачу побудови мережі предметних областей і дерева понять [4] на базі заданого концепту. Зображені у статті підходи розширюють шляхи інтерпретації отриманих результатів за рахунок апробації на окремому концепті.

Виклад основного матеріалу дослідження. 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:

1. Computer Science (41 піднапрямок)
2. Economics (1 піднапрямок)
3. Electrical Engineering and System Science (3 піднапрямки)
4. Mathematics (32 піднапрямки)
5. Physics (51 піднапрямок)
6. Quantitative Biology (10 піднапрямків)
7. Quantitative Finance (10 піднапрямків)
8. Statistics (6 піднапрямків)

Для кожного наукового напряму було сформовано словник, який містить назву наукового напряму і має вигляд таблиці (табл. 1), що складається з трьох полів із такою інформацією:

Таблиця 1

Частина словника «Computer Science»

№ 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 р. ресурсом було запроваджено новий пошуковий інтерфейс, який передбачає скорочене зазначення назви піднапрямку (рис. 1) в переліках результатів пошуку, тож, на відміну від попереднього представлення результатів, виникла необхідність тлумачення скороченої назви піднапрямку для подальшої коректної візуалізації результатів.

Робота розробленої системи передбачає задання концепту для пошуку та, за результатами пошуку, виокремлення наукових напрямків і піднапрямків, у рамках яких було здійснено дослі-


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

Рис. 1. Сторінка результатів пошуку на ресурсі препринтів arXiv.

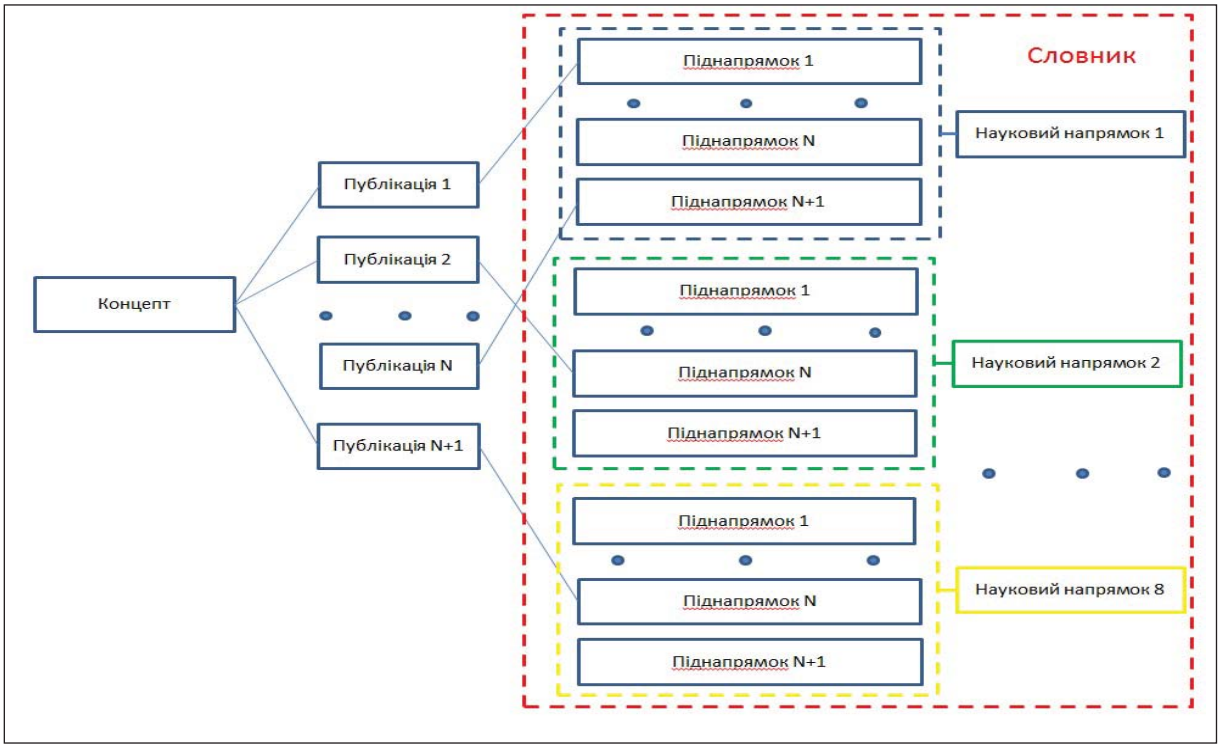

Рис. 2. Модель «Концепт-система наукових напрямків»

дження й опубліковано результати на ресурсі, і подальше представлення результатів.

Також для масиву публікацій було застосовано додатковий пошук, що виокремлює публікації, які були опубліковані в наукових виданнях.

Модель «Концепт – система наукових напрямків» можна представити у вигляді схеми (рис. 2), що становить масив наукових напрямків і піднапрямків, які формують словник і безпосередньо зв'язок концепту із цими параметрами.

Алгоритм проведення дослідження представлений на рис. 3.

Для реалізації поставленої задачі було застосовано такі методи:

1. Метод аналізу текстових масивів, що застосовується для виокремлення з тексту інформації щодо наукового піднапрямку, до якого віднесено публікацію, а також виокремлення публікацій, що вже опубліковані в наукових виданнях.
2. Методи математичної лінгвістики – формування словників та оцінка текстового пошуку і співставлення.
3. Методи статистичного аналізу, що дозволяє провести підрахунки співвідношення кількості публікацій – відповідно піднапрямків і напрямків.
4. За теорією графів було побудовано візуалізацію результатів роботи.

Рис. 3. Узагальнений алгоритм збору та візуалізації отриманої інформації

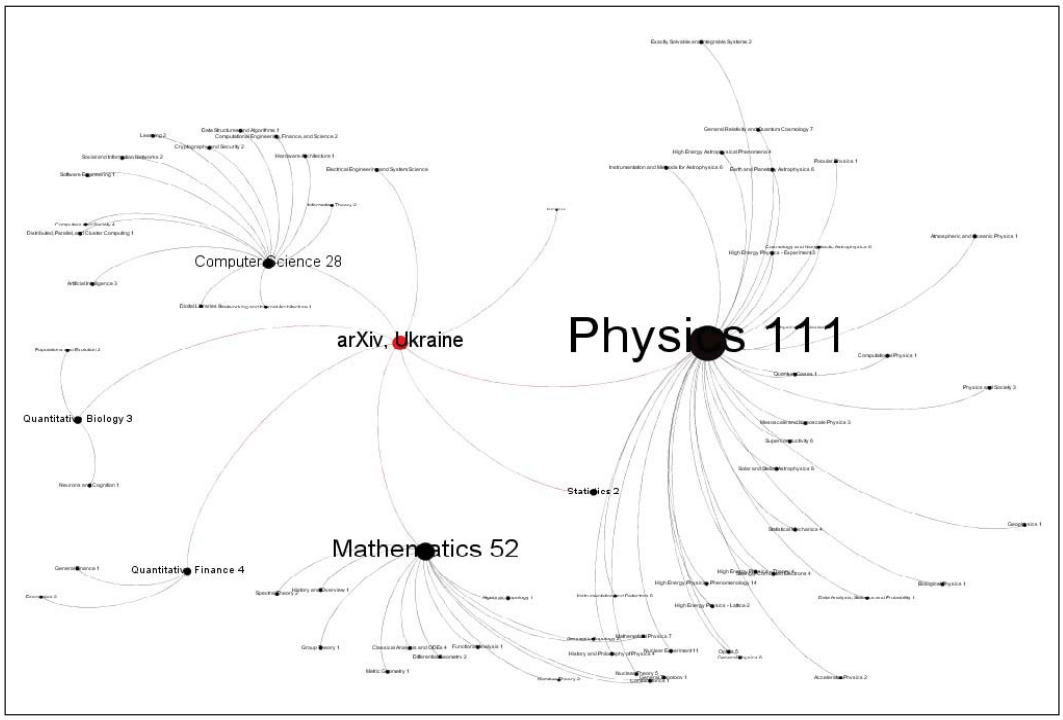

Рис. 4. Візуалізація результатів пошуку публікацій за заданим концептом на ресурсі arXiv

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

Алгоритм було реалізовано мовою програмування Java в середовищі Eclipse.

За заданим концептом було виокремлено 619 публікацій. До уваги було взято 200 публікацій за період 2009–2018 рр. 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 були опубліковані та візуалізовані в середовищі Gephi.

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

1. Physics – 111 публікацій за 32 піднапрямками.
2. Mathematics – 52 публікації за 13 піднапрямками.
3. Computer Science – 28 публікацій за 13 піднапрямками.
4. Quantitative Finance – 4 публікації за 2 піднапрямками.
5. Quantitative Biology – 3 публікації за 2 піднапрямками.
6. 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 такі:

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

Вищезазначені результати було візуалізовано у вигляді ненаправленого графа (рис. 3, 4). Вершинами графа є наукові напрями і піднапрямки із зазначенням кількості публікацій за заданим концептом, розміщених на ресурсі препринтів.

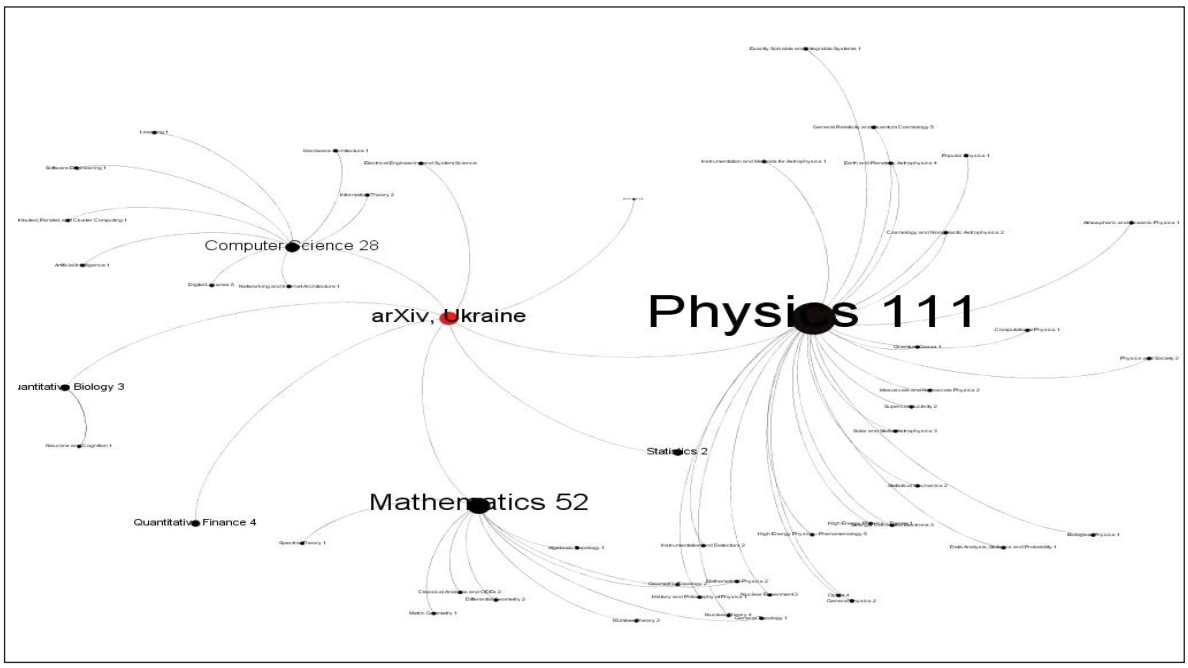

Рис. 5. Візуалізація результатів пошуку публікацій за заданим концептом, що були опубліковані в наукових журналах і матеріалах конференцій.

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.

Візуалізація результатів пошуку препринтів за окремою схемою – тих препринтів, що були опубліковані – дає можливість оцінити, яка частка розміщених матеріалів надалі була опублікована.

Висновки. У роботі було запропоновано модель, створену на базі ресурсу препринтів arXiv «Концепт – система наукових напрямків». Також було зображено порядок роботи системи з пошуку й аналізу інформації за результатами пошуку за заданим концептом. Обраний для апробації концепт та обсяг результатів пошуку дозволив перевірити працездатність розробленого програмного забезпечення. Результати пошуку було побудовано за такими характеристиками публікації, як науковий піднапрямок, наявність опублікованого матеріалу. Подальший розвиток моделі за рахунок врахування таких елементів, як автор, назва наукового видання, ключові слова, може зробити результати пошуку більш інформативними.

Сьогодні розроблена система може виступати допоміжним інструментом в оцінці результативності досліджень, порівняльному аналізі та прийнятті рішень щодо надання підтримки науковим проектам для розвитку за результатами дослідження.

Список літератури:

1. Sorokina D. Plagiarism Detection in arXiv. URL: https://ieeexplore.ieee.org/abstract/document/4053155/.
2. Warner S. Open Archives Initiative protocol development and implementation at arXiv. URL: https://arxiv.org/abs/cs/0101027.
3. Юрченко О.В. Дефініція концепту в сучасних лінгвістичних дослідженнях.URL: http://web.znu.edu.ua/herald/issues/2008/fil_2008_1_2/2008-26-06/yurch.pdf.
4. Андрущенко В.Б. Побудова дерева предметних областей для заданого поняття на базі ресурсу препринтів ArXiv. Матеріали XI Міжнародної науково-технічної конференції «Інтелектуальні технології лінгвістичного аналізу», 24–25 жовтня 2017 р.; Національний авіаційний університет. Київ: НАУ, 2017. С. 20.

НОВЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ ПОИСКА, ОБРАБОТКИ И ИНТЕРПРЕТАЦИИ ДАННЫХ РЕСУРСА ARXIV

Статья посвящена новым технологиям работы с ресурсом открытого доступа – архивом препринтов arXiv. За счет представления данных, которые содержит ресурс в виде модели «Концепт – система научных направлений», разработан и представлен алгоритм работы с ресурсом и выделено новые массивы информации, которые были сформированы с применением методов text mining к результатам поиска на ресурсе по заданному концепту. В статье показаны результаты апробации алгоритма, их визуализация в среде Gephi и предложены подходы к интерпретации полученных данных. Также были представлены направления развития реализованной задачи.

Ключевые слова: архив препринтов, концепт, научное направление, публикация, граф.

NEW INFORMATION TECHNOLOGIES FOR ARXIV DATA SEARCH, PROCESSING AND ANALYSIS

Paper is dedicated to representing of new technologies for arXiv data search, processing and analysis. By presenting data of archive as a model "Concept – System of research fields" there was developed and realized the algorithm of arXiv data processing and there were obtained new arrays of information formed by application of text mining methods to the search results for the concept. The results of algorithm testing results, visualization in Gephi are depicted in paper and also there suggested approaches to obtained data interpretation and ways to develop the task.

Key words: *preprints archive, concept, research field, publication, graph.*