УДК 004.4

Д. Тарасов, О. Гарасим
Національний університет “Львівська політехніка”,
кафедра СКІД

ВИКОРИСТАННЯ СПЕЦІАЛІЗОВАНИХ ПОШУКОВИХ СИСТЕМ ДЛЯ ОТРИМАННЯ ПОКАЗНИКІВ ЦИТОВАНОСТІ ЕЛЕКТРОННИХ НАУКОВИХ АРХІВІВ

Розглянуто проблему визначення індексу цитованості документів електронного наукового архіву та обчислення наукометричних показників українських науковців. Розроблено алгоритм пошуку цитованості для електронного наукового архіву. Побудовано прототип інформаційної системи аналізу наукометричних показників електронних наукових документів.

Ключові слова: індекс цитованості, наукометрія, електронний наукових архів, Google Академія.

In the paper raised the problem of searching documents index citation in electronic scientific archive and computing scientometrics indicators of Ukrainian scientists. Built algorithm of citation searching for electronic scientific archive. Built a prototype information system analysis of scientific metrics of scientific documents.

Key words: citation index, scientometrics, electronic scientific archive, Google Scholar.

Вступ. Загальна постановка проблеми

Останніми роками необхідним є використання індексу цитованості для відображення рівня наукових досліджень окремого автора чи наукових товариств. У суспільстві цікавляться результатом, який надає робота, а не процес діяльності чи думки. Із збільшенням кількості науковців виявляється питання про науковий внесок кожного з них. Для порівняння необхідні кількісні показники, що пов’язано із безліччю проблем, основні з яких пов’язані із врахуванням якісного характеру наукової роботи, а, по-друге, інтерпретування показників у числовій розмірності. Вирішуючи такі проблеми, можна аналізуванням набути знання про актуальність певної тематики і навпаки про застарілість, рівень опису сучасних проблем тощо.

Доволі тривалий час показником наукового рівня була кількість публікацій, але вона зовсім не відображає корисність суспільству. З 2005 р. з’являється поняття “індекс імені”, який обчислюється на основі розподілення цитувань робіт. Діяльність з розроблення нових методів оцінення корисності наукових публікацій та модернізацій наявних є доволі осмислена і затребувана.

Інформаційні ресурси потребують кількісного вивчення потоків інформації, зокрема – наукової інформації. У такій складній та важко формалізованій системі, як наука саме інформація є основним ресурсом та продуктом одночасно. Тому розвиток методів аналізу потоків наукової інформації дозволяє наблизитись до розуміння функціонування та структури науки. Зокрема, важливо також зрозуміти, за якими принципами здійснюється пошук та використання інформації самими учасниками наукового процесу. Завдяки можливості збору та обробки значних обсягів статистичних даних, почали розвиватись порівняно нові напрямки вивчення інтелектуальної структури та характерних ознак розвитку науки – інформетрія (infometrics) [1], бібліометрія (bibliometrics) [2], наукометрія (scientometrics) [3], вебометрія (webometrics) [4].

Зв'язок висвітленої проблеми із важливими науковими та практичними завданнями

З розвитком єдиного інформаційного простору та інтенсивним впровадженням інформаційно-комунікаційних технологій призвели до стрімкого накопичення документів та інформаційних ресурсів, значну частку яких становлять наукові архіви.

Існують засоби визначення індексу цитованості: Scopus, Web of Science, Російський індекс наукового цитування, основою яких є комерційні дані. Великий обсяг інформації для застосування наукометричних та бібліометричних технологій стосовно праць українських вчених зберігаються у відкритих електронних архівах [14] та не аналізуються згаданими системами пошуку індексу цитування.

Актуальність теми зумовлена стрімким розвитком наукової діяльності, яку важко оцінити лише одним параметром, сьогодні усе більшою є необхідність оцінювання з використанням кількісних параметрів, що характеризують наукову діяльність. Показники індекс цитування та імпакт-фактор для оцінювання результатів наукових досліджень і наукових періодичних видань використовують для зіставлення рівня наукових досліджень різних організацій чи окремих дослідників.

Мета і задачі дослідження. Побудувати БД інформаційної системи, де розміщуватимуться формалізовані дані з електронного наукового архіву та отримані дані з пошукової системи наукового індекса цитованості про архівний документ. Для досягнення мети були поставленні такі задачі:

- вибрати кращу систему пошуку статей, для здійснення індексування наукового електронного ресурсу;
- розробити процедуру отримання показника цитованості праць ВНЗ з використанням інструментів ЕНА та інтернет пошуковими системами;
- реалізувати модуль для автоматизації пошуку індексу цитованості.

Аналіз останніх досліджень та публікацій

Індекс цитованості є доволі важливим, який можна отримати, використовуючи сучасні пошукові системи, що спеціалізуються на індексації наукової літератури. Індекс цитування – прийнята в науковому світі міра значущості наукової роботи будь-якого вченого або наукового колективу. Величина індексу цитування визначається кількістю посилань на публікацію або прізвище автора в інших джерелах [5]. Для того, щоб здійснити пошук цитованості наукових документів електронного архіву, розглянемо характеристики систем наукової індексації, основними з яких є Scopus, Російський індекс наукового цитування, Web of Science (WoS), Google Scholar.

Scopus являє собою найбільшу в світі єдину реферативну базу даних, яка індексує більше ніж 17000 назв науково-технічних і медичних журналів приблизно 4000 міжнародних видавництв. Щодня оновлювана база даних Scopus містить записи до першого тому, першого випуску журналів провідних наукових видавництв. Вона забезпечує неперевершену підтримку в пошуку наукових публікацій і пропонує посилання на всі цитати з широкого обсягу доступних статей. База даних Scopus охоплює близько 25048 видань, але серед них є тільки 35 українських видань станом на 2010, тобто 0.15 %. При цьому з 35 лише 18 активних видань, що постійно оновлюють інформацію [6, 7].

Російський індекс наукового цитування (РІНЦ) спочатку був задуманий як база даних і пошукова система, здатний забезпечити повноцінне покриття російського наукового простору. Зробити це можливо, лише використовуючи потужності IT-технологій. Несподівано виявилося, що середня кількість бібліографічних посилань на друковану статтю – 2,74; середня кількість посилань на безкоштовну електронну версію – 7,3. Тобто в 2,6 раза більше ніж на друковану. Електронні статті цитують у 4,5 раза частіше від друкованих і це співвідношення швидко зростає [8].

Google Академія є вільно доступною пошуковою системою, яка індексує повний текст наукових публікацій всіх форматів і дисциплін. Індекс Google Scholar містить більшість рецензованих онлайн-журналів Європи та Америки найбільших наукових видавництв [9]. Пошукова програма Google Scholar працює за тими ж правилами, що і пошукова програма Google, але індекси Google і Google Scholar – це різні бази даних. Google Scholar виконує не тільки інформаційні, але й наукометричні функції [10].

Web of Science – база даних, яку публікує Філадельфійський інститут наукової інформації, визначаючи журнали з найвищим рейтингом (Thomson Reuter Master Journal List). Для визначення індекса цитованості вченого використовуються БД Science Citation Index Expanded і Social Sciences Citation Index. У WoS закладена можливість пошуку процитованих робіт не тільки по першому автору, а й по співавторах, за умови, що джерело, в якому міститься стаття, розписаний в Web of Science. Можна задати ретроспективу індексу цитованості і хронологічні межі процитованих робіт [11].

Microsoft Academic Search є вільною академічною пошуковою системою наукових досліджень, розробленикомякої є Microsoft. Вона охоплює понад 36 мільйонів публікацій і більше 18 мільйонів авторів у різних галузях науки, оновлення додається кожного тижня. [15]

У джерелі [12] здійснений порівняльний аналіз систем пошуку індексу цитованості.

PubMed і Google Scholar вільні і забезпечують відкритий доступ для всіх зацікавлених користувачів. Scopus і WoS є базами даних, які належать комерційним провайдерам і є платними.

PubMed орієнтований переважно на медицину та біомедичних наук, у той час як Scopus, Web наук і Академія Google охоплюють більшість галузей науки. WoS охоплює найстаріші видання. PubMed здійснює пошук за більшою кількістю ключових слів. Scopus містить статті, опубліковані з 1966 р., але інформація про цитатуваний аналіз доступна тільки для статей, опублікованих після 1996 р. На відміну від PubMed, Scopus і WoS легко оновлені для друкованої літератури, але не містить старіші видання.

Індекс цитатування в Scopus представлений у вигляді таблиці з номерами цитованих статей за окремі роки, а також загальну кількість цитованих посилань за всі роки. Цитованість можна отримати, просто натиснувши на кількість посилань.

Визначимо, що таке цитування: (1) стаття А цитує статтю Б, якщо хоч би один раз в тексті А є посилання на Б і Б, таким чином, винесена в А в пристатейний список літератури або фігурує в посторінковій виносці; (2) журнал М цитує журнал N так: скільки статей з М цитують статті з N. Таким чином, якщо в тексті однієї статті інша публікація згадується кілька разів, це вважається одним цитуванням.

У джерелі [13] подані основні методи одержання числових показників цитованості:

"Класичний" (синхронний, Гарфильдівський) імпакт-фактор. Класичний імпакт-фактор, тобто те, що розуміють під ним за замовчуванням, – це, в строгому визначенні, "синхронний дворічний імпакт-фактор, без урахування поточного року". Саме він обчислюється Інститутом наукової інформації ISI і щорічно публікується в базі даних Journal Citation Reports. Саме він в наш час фігурує при порівнянні рівня журналів. Класичний імпакт-фактор журналу визначається відношенням кількості журналів, що з'явилися в усьому масиві, за рік Y посилань на статті журналу М, що вийшли в роках Y – 1 та Y – 2, до сумарної кількості статей, що вийшли в М за той же період, роках Y – 1 і Y – 2.

"Диахронний" імпакт-фактор. Диахронний імпакт-фактор журналу М в році Y задається відношенням числа посилань з журналів за рокиY+2 і Y+1 на статті журналу М, що вийшли в році Y, до сумарного числа статей, що вийшли в М в році Y.

Зважаючи на те, що у разі синхронного імпакт-фактора фіксується рік цитування і досліджується, які статті процитовані з минулих років, а у разі диахронного – фіксується рік публікації і підраховуються майбутні цитування опублікованих в цьому року робіт, синхронний підхід також називають ретроспективним, а диахронний – перспективним.

Індекс Хірша (h–індекс) – наукометричних показник, запропонований у 2005 американським фізиком Хорхе Хірш з університету Сан-Дієго, Каліфорнія. Індекс Хірша є кількісною характеристикою продуктивності вченого, заснованої на кількості його публікацій і кількості цитувань цих публікацій. Індекс обчислюється на основі розподілу цитувань робіт даного дослідника. Вчений має індекс h, якщо h з його Np статей цитуються як мінімум h раз кожна, в той час як решта (Np – h) статей цитуються не більше ніж h раз кожна.

Основні функції цитованості:

- інформаційний пошук для обслуговування індивідуальних дослідників і наукових організацій;

- використання зв'язків між публікаціями для виявлення структури галузей знання, спостереження і прогнозування їх розвитку (створюється “карта” науки і виявлення дослідницьких фронтів);
- оцінка якості публікацій і їх авторів науковим співтовариством.

Комплексна оцінка індексу цитування дозволяє оцінювати наукові підрозділи по науковцях, що до них входять. Вона широко використовується для оцінювання журналів, наукових товариств, редакційних колегій тощо.

Виділення проблеми

2 грудня 2009 р. у науково-технічній бібліотеці Львівської політехніки відкритий науковий електронний архів (http://ena.lp.edu.ua), що базується на платформі DSpace компанії Hewlett-Packard, основними перевагами якого є швидкий пошук, завантаження, перегляд електронних наукових документів. У електронний архів наукових публікацій НТБ увійшли науково-технічні збірники, журнали, матеріали конференцій, публікації та електронні документи працівників НТБ, бібліографічні видання та багато іншого.

Науковий електронний архів (ЕНА) Національного університету “Львівська політехніка” виконує такі основні функції:

- приймання – процеси отримання, перевіряння надісланого архівного документа в електронному вигляді на відповідність визначеним вимогам до носіїв інформації, форматів даних, необхідної довідкової інформації, а також того, чи профільний документ є цілісним та автентичним;
- архівне зберігання – процеси архівного описування, розміщення прийнятого архівного документа в електронному вигляді на носієві інформації для зберігання у відповідності з прийнятою ієрархією; оновлення носіїв інформації згідно з планом здійснення процесів зберігання; періодичного перевіряння цілісності профільних документів, резервного та страхового копіювання; забезпечення доступу до профільних документів за вимогою;
- керування даними – керування процесами функціонування електронного архіву у відповідності з визначеним планом зберігання та адміністративними заходами, спрямованими на керування процесами приймання, зберігання та доступу до документів в електронному вигляді;
- організація доступу – забезпечення користувачів та державні архіви інформацією щодо наявності електронного документа на зберігання та реалізації отримання їх за запитом через телекомунікаційні засоби відповідно до процедур, що обмежують доступ та забезпечують захист інформації;
- планування процесу – функція планування управління процесами електронного архіву відповідно до запланованих подій життєвого циклу архівних документів та програмно–технічних засобів;
- адміністрування – функція керування процесами електронного архіву, що відповідає вимогам поточної експлуатації електронного архіву;
- організація механізм співпраці учасників проекту (бібліотекарів та вчених) в цілях поширення наукової інформації, популяризувати науково-дослідницьку роботу в Університеті.

ЕНА діє на системі DSpace. DSpace – це програмне забезпечення для створення архіву електронних ресурсів (цифрового сховища). Платформа DSpace розроблялася спільно компанією Hewlett-Packard і бібліотеками MIT (Massachusetts Institute of Technology). Четвертого листопада 2002 система була запущена як діюча служба, підтримувана бібліотеками MIT. Також на підставі ліцензії BSD відкритий вихідний код з наміром заохотити формування спільноти відкритих кодів навколо DSpace.

DSpace – це безкоштовна і проста в установці система і повністю настроюється під потреби будь-якої організації, вона зберігає всі типи цифрового контенту, включаючи текст, зображення, динамічні зображення, MPEG тощо [5].

Враховуючи сучасні тенденції до необхідності визначення індексу цитованості наукових робіт авторів, постає проблема його визначення для українських вчених. Проблема полягає в тому, що значна частина опублікованих праць не входять до бази аналізу систем пошуку індексу цитованості. Лише Google Scholar дає можливості такого аналізу для українських вчених, завдяки індексуванню відкритих електронних архівів створених українським навчальним закладом.

Google Scholar ставить за мету узагальнити всі електронні посилання по темі. Існує список джерел, доступних для Google Scholar, тому що вона містить список всіх публікацій, які вийшли з електронного пошуку. Будучи по суті web-пошуком, його мета полягає в досягненні широкої аудиторії. Це дозволяє здійснювати швидкий і розширений пошук. У розширеному пошуку результати можуть бути обмежені за назвою слів, авторів, джерела, дати публікації, предметних областей. Мова інтерфейсу та пошуку є необов'язковою. Результати можуть бути відображені як список 10–300 елементів на сторінці. Кожна отримана стаття містить назву, прізвища авторів і джерело. Під кожною статтею вказується число цитувань, якщо воно існує, а також можна перейти по лінку, де вказані джерела, які посилаються на статтю.

Ще однією проблемою є отримання відомостей про документи, які проіндексовані пошуковою системою, а також отримання показників цитованості для того, щоб продовжити подальші дослідження ЕНА. Для вирішення цієї проблеми необхідно розробити модуль для здійснення автоматизованого пошуку та отримання індексу цитування.

Аналіз отриманих результатів

Електронний науковий архів (ЕНА) відкриває можливості дослідження наукового внеску зареєстрованих дослідників, розвиток окремої галузі науки, наукового видання, публікації, а також загальний науковий рівень університету. Сьогодні в архіві зберігаються понад 10000 наукових праць, з яких 1930 проіндексовані Google Академією і 153 публікації мають індекс цитування. Науковий внесок електронних документів оцінюватимемо, шляхом пошуку індексу цитування в базі даних Google Scholar.

Враховуючи переваги Google Scholar, до яких зараховуємо: активне індексування наукових документів, доступність, зручність використання, індексування українських наукових баз даних, вибираємо його для подальших досліджень наукового потенціалу Електронного наукового архіву.

Основними елементами розробленого модуля, який забезпечує пошук та отримання індексу цитування наукових документів ЕНА від Google Scholar, є (див. рис. 1):

- “Опрацювання даних” – метою якого є відправлення запитів до ЕНА для отримання необхідних даних та адаптувати їх для можливості подальшої роботи з ними. Вибирає тільки необхідні дані, містить правила для коректного пошуку, фільтрації даних;

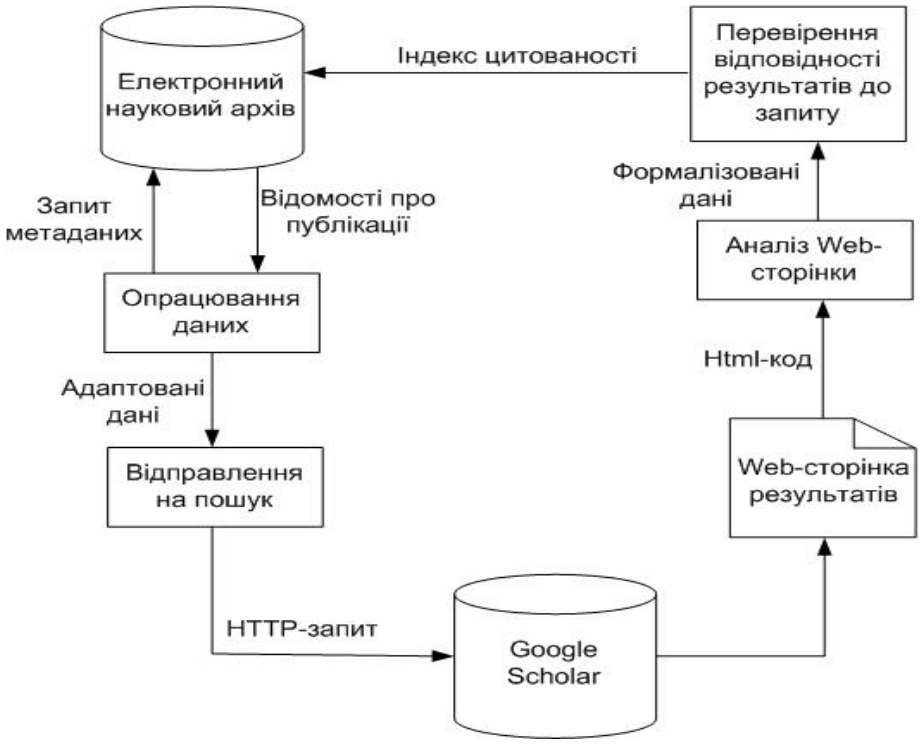

Рис. 1. Процес отримання індексу цитованості від Google scholar для ЕНА

• “Відсилання на пошук” – задає правила пошуку, містить налаштування для точнішого пошуку (пошук у визначеному домені, вибір галузі науки, можливості простого та розширеного пошуку). З’єднується з базою даних Google Scholar та відсилає http-запит;

• “Google Scholar” – система визначена як така, що найкраще задовольняє вимоги пошуку індексу цитованості наукових документів українських дослідників. Можна використовувати іншу систему;

• “Web-сторінка результатів” – містить результати пошуку, передається у вигляді html-коду;

• “Аналіз Web-сторінки” – проводиться розбір результатів та їх формалізація;

• “Перевірення відповідності результатів до запиту” – містить правила перевірення на відповідність отриманих результатів науковому документу. Результати записуються до ІС.

Схема бази даних ІС показана на рис. 2, яка є прототипом для подальших досліджень. Будемо її використовувати для тестування вирішення проблеми: визначення індексів цитованості, наукометричних показників наукових документів електронного архіву.

База даних складається з восьми таблиць: DC_Publication, DT_Publ_Citation, RF_Journal, RF_Journal_Edition, які містять відомості про електронний науковий документ, результати пошуку індексу цитованості, вихідні дані про видання.

DC_Publication – сутність, яка відображає первинну інформацію, яка отримана і адаптована з ЕНА. Для опису кожної публікації ЕНА використовуємо атрибути: Title (назва), Date_year (рік опублікування), Original_date (рік створення), Collaborators (співавтори), Issue_ID (номер видання), Type_publication (тип публікації), Handle (унікальний номер наукової публікації, який є її ідентифікатором в ЕНА).

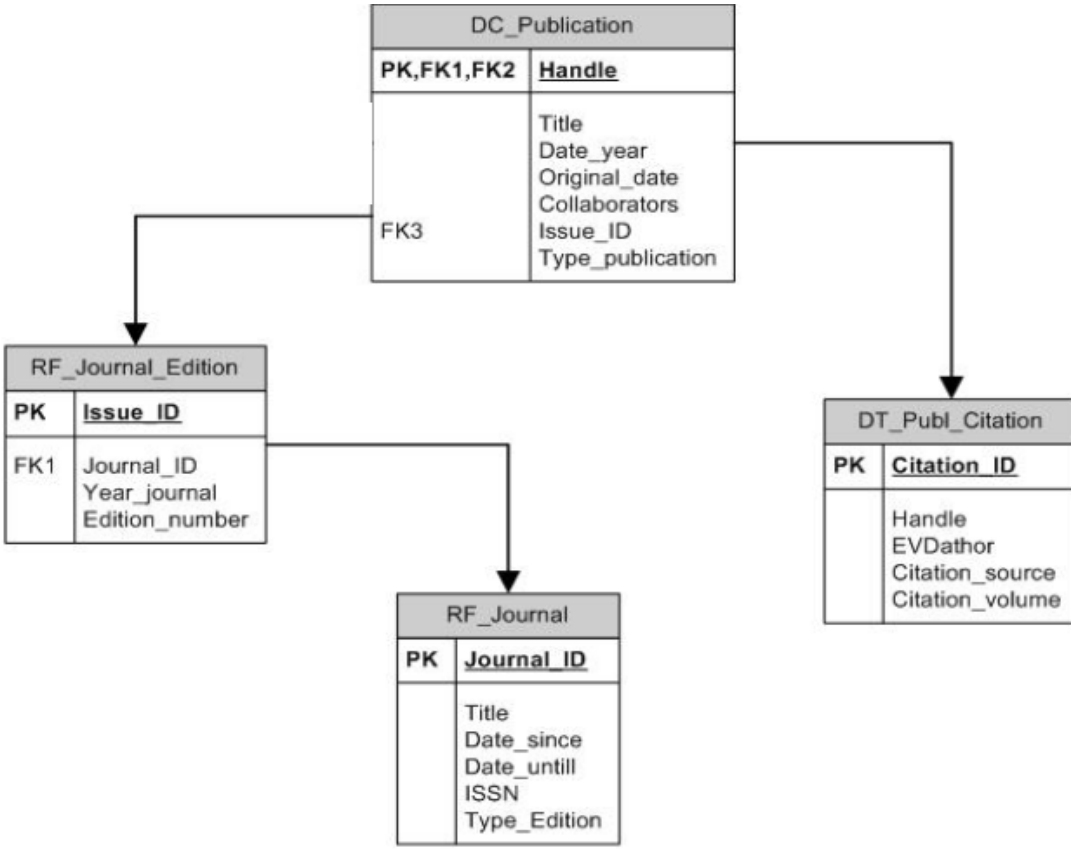

Рис. 2. Схема бази даних ІС

RF_Journal_Edition – містить відомості про видання. Атрибутами, які описують сутність є: Issue_ID (номер-ідентифікатор видання), Journal_ID (оригінальний номер видання), Edition_number (номер випуску), Year_journal (рік випуску).

RF_Journal – описує періодику видання. Title (назва), Date_since (дата початку випуску видання), Date_until (дата закінчення випуску видання), ISSN, Type_Edition (тип публікації).

DT_Publ_Citation – сутність, яка описує знайдені результати пошуку модуля. Citation_ID (номер-ідентифікатор запису), Handle (унікальний номер наукової публікації, який є її ідентифікатором в ЕНА), EVDathor (дата здійснення пошуку), Citation_source (назва пошукової системи), Citation_volume (число цитованості електронного документа).

Висновки і перспективи подальших наукових розвідок

Українські університети відсутні в міжнародних дослідницьких рейтингах наукової діяльності, що обмежує українських вчених у підтверджені вагомості своєї наукової роботи. Хоча в Україні відбувається розгортання системи моніторингу наукового потенціалу, але це відбувається доволі повільними темпами. Отже, подальші дослідження будуть спрямовані на математичне обчислення наукометричних показників електронного наукового архіву університету, використовуючи вільно вживану систему пошуку індексу цитування.

Розроблено модуль, який отримує індекс цитованості з Google Scholar, а також враховує обмеження і передбачає можливі помилки виданих результатів Google Scholar. Модуль дозволяє здійснювати пошук за доменом та фільтрувати дані, що підвищує адекватність оцінки електронних документів архіву.

У результаті аналізу систем пошуку індексу цитованості виділено Google Scholar, основною перевагою якої є опрацювання українських видань. Для досліджень використовується індекс цитованості наукових публікацій з вільно вживаної системи аналізу наукових робіт. Побудовано прототип інформаційної системи аналізу наукометричних показників електронних наукових документів, яка складається з бази даних, де зберігаються дані про електронні наукові документи та модуля, який здійснює видобування, формалізування, фільтрування даних, накладає обмеження та отримує результати. Розроблено алгоритм пошуку цитованості для електронного наукового архіву.

1. Egghe L. Introduction to Informetrics : quantitative methods in library, documentation and information science / Egghe L., Rousseau R. // Elsevier Science Publishers. – Belgium, 1990. – 327 c. 2. Glänzel W. Bibliometrics as a research field [Електронний ресурс]. – Режим доступу: http://nsdl.niscair.res.in/bitstream/123456789/968/1/Bib_Module_KUL.pdf 3. Хайтун С.Д. Наукометрия. Состояние и перспективы / Хайтун С.Д. // Наука: М, 1983. – 183 с. 4. Almind T. Informetric analyses on the World Wide Web: Methodological approaches to "webometrics" / Almind T., Ingwersen P. // Journal of Documentation, 1997. – № 53 (4). – P. 404–420. 5. Показники цитування [Електронний ресурс]. – Режим доступу: http://ntb.pstu.edu/ index.php?id=22 6. Scopus [Електронний ресурс]. – Режим доступу: http://ntb.pstu.edu/index.php?id=59&L=0 7. Jacso P. As we may search – Comparison of major features of the Web of Science, Scopus and Google Scholar citation-based and citation-enhanced databases // Current Science – 2005 – V.89, N 9, 10. – P. 1537 – 1547 8. РІНЦ [Електронний ресурс]. – Режим доступу: http://ntb.pstu.edu/index.php?id=61&L=0 9. [Електронний ресурс]. – Режим доступу: http://articles.tutorialonline.biz/portal/language-ru/Google%20Scholar 10. Google Scholar [Електронний ресурс]. – Режим доступу: http://www.abc.chemistry.bsu.by/intro/part10/04.html 11. Рекомендации относительно поиска журналов, которые входят в международные научно-метрические базы данных (SCOPUS, Web of Science) [Електронний ресурс]. – Режим доступу: http://ru.snu.edu.ua/index.php?mode=392 12. Falagas M.E. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses / Falagas M.E., Pitsouni E.I., Malietzis G.A., Pappas G. // FASEB J. – 2008. 13. Проблеми оцінки наукової діяльності [Електронний ресурс]. – Режим доступу: http://webometr.kpi.ua/node/53 14. Тарасов Д.О. Технологічні особливості редагування інформації у електронних архівах / А.І. Андрухів, Д.О. Тарасов // Матеріали другої науково-практичної конференції "Сучасні проблеми діяльності бібліотеки в умовах інформаційного суспільства ", Львів, 23 вересня 2010 р. / Нац. ун-ет "Львівська політехніка", Науково-технічна бібліотека . – Львів: Вид-во Нац. ун-ту "Львівська політехніка", 2010. – С. 99–104. 15. Microsoft Academic Search [Електронний ресурс]. – Режим доступу:http://academic.research. microsoft.com/About/Help.htm#1