УДК 51.001.57

Д.О. ТАРАСОВ, канд. техн. наук, доц., НУ "Львівська політехніка", Львів,
О.Р. ГАРАСИМ, асп., НУ "Львівська політехніка", Львів.

АЛГОРИТМ ОТРИМАННЯ ДАНИХ ДЛЯ ПОБУДОВИ ІНДЕКСУ ЦИТОВАНОСТІ ДОКУМЕНТІВ ВІДКРИТИХ ЕЛЕКТРОННИХ АРХІВІВ

Обгрунтовано необхідність створення алгоритму для отримання цитованості наукових документів відкритих електронних архівів. Побудований алгоритм отримання індексу цитованості з спеціалізованої системи Google Scholar. Описані основні кроки роботи алгоритму Іл.: 1. Бібліогр.: 8 назв.

Ключевые слова: Google Scholar, індекс цитованості, алгоритм отримання індексу цитованості.

Постановка проблеми. Відкриті електронні архіви України швидко розвиваються, щодня поновлюються новими науковими документами, а отже можуть надавати оцінку окремим науковцям, виданням, журналам. Для отримання індексу цитованості необхідний алгоритм, який би регулярно в автоматизованому режимі здійснював перевірки цитувань наукових документів, будував рейтинги та статистичні графіки змін в електронному архіві. Головна проблема при реалізації алгоритму полягає в тому, що необхідно адаптувати дані відкритого електронного архіву для точного пошуку документа, враховуючи обмеження спеціалізованої системи Google Scholar.

Аналіз літератури. Сьогодні електронні бібліотеки та архіви розширюються [1], щодня поповнюючи свої колекції новими документами. Для того, щоб забезпечити довготривале існування та інтенсивне оновлення даних електронного архіву, постає необхідність у задоволені нових вимог користувачів. Користувачі електронних бібліотек і архівів прагнуть до швидкого пошуку матеріалів, зручного інтерфейсу, можливості створення власної сторінки із збереженими документами та відслідковувати частоту їх перегляду відвідувачами [2].

Під індексом цитованості розуміють відстеження посилань, які автори записують в бібліографію опублікованих робіт, що відображається в кількісному значенні. Індекс цитованості дає можливість для пошуку і аналізу релевантної та актуальної літератури. Він також дозволяє користувачам збирати дані про влив журналів, а також оцінити по конкретних предметних областях наукову діяльність [3].

Індекс цитованості в науці використовують для:

1. Пошуку документів, які наводять на більш ранні документи. Індекс цитованості – це спосіб оглянути розвиток наукової думки від початкового документа. Тобто це зручний спосіб пошуку наукових матеріалів.

2. Отримання показників цитування власних публікацій.

3. Визначення рейтингу популярності журналів в галузі. Цитування використовується для ранжирування журналів, зокрема за предметними областями, як правило, на основі імпакт-фактора.

4. Перевірення старих цитувань. Часто посилання здаються незрозумілими, але коли пройти за пошуковим "ланцюжком" можна отримати повне відображення звернень на власну публікацію.

Індекс цитованості також використовують і в бізнесі. Часто наукові дослідження стають проривом у техніці чи технології, відкривають шляхи розв'язання важких ситуацій. Але будь-які дослідження потребують інвесторів, яких потрібно переконати у потрібності виконання робіт. Сьогодні часто звертають увагу на індекс цитування при прийнятті рішень про майбутні інвестиції [4].

Основною перевагою Google Scholar є вільний доступ до інтелектуальних ресурсів дослідних закладів. Google Scholar має унікальну навігацію в Інтернеті, надаючи можливість знайти пов'язані статті, не обмежуючись назвою, мовою, територією чи авторами. Це спеціалізована система, яка охоплює практично всі науки і дисципліни [5, 6].

Метрики цитування [7]:

– загальна кількість документів;
– загальна кількість цитат;
– середнє число посилань на документ;
– середнє число посилань на автора;
– середня кількість робіт на одного автора;
– *h*-індекс;
– *g*-індекс.

Ціль статі – розроблення алгоритму отримання індексу цитованості документів відкритих електронних документів з спеціалізованої системи Google Scholar.

Алгоритм отримання індексу цитованості наукових документів. В джерелі [8] подані технічні інструкції з підвищення точності запитів, що враховано при проектуванні алгоритму та доповнено ще кількома моментами для абсолютно точного пошуку (див. рис.).

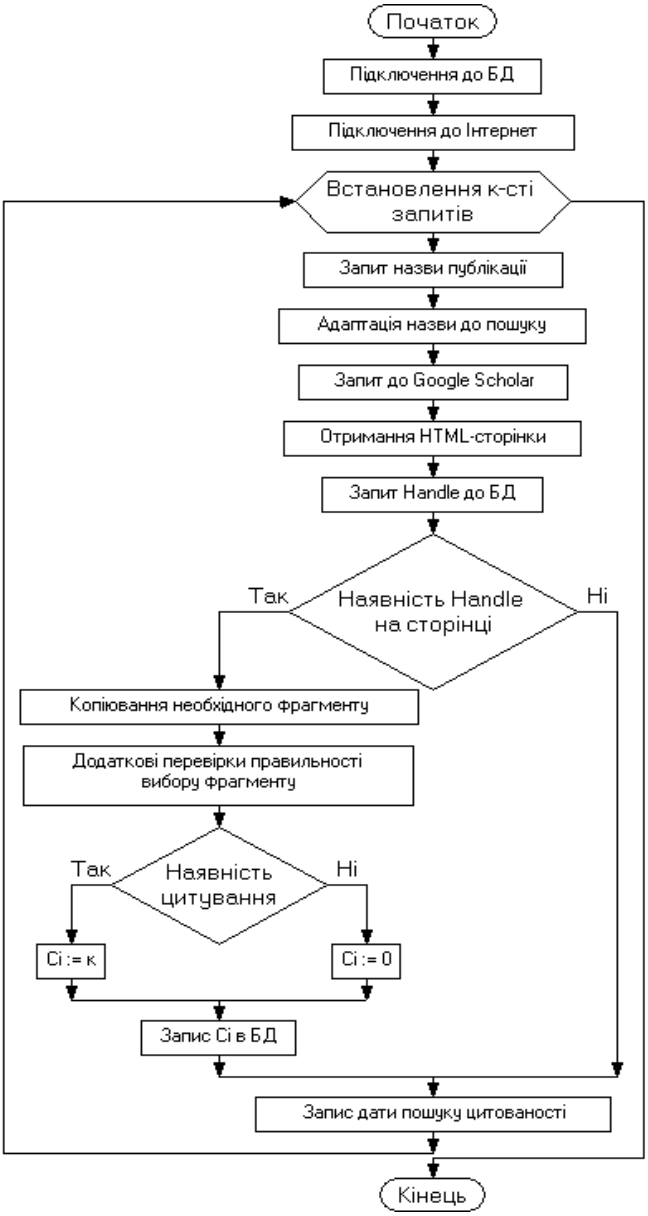

Рис. Алгоритм отримання індексу цитованості наукових документів електронного архіву

Алгоритм отримання індексу цитованості наукового документа електронного архіву починається з підключення до БД інформаційної системи та Інтернет.

Наступний крок є дуже важливий – "Встановлення к-сті запитів", оскільки Google Scholar блокує активні запити з метою недопущення "зависання" сторінки. З метою дотримання прав Google встановлюється число запитів, після яких на певний період запити припиняються.

"Запит назви публікації" – крок на якому з БД інформаційної системи вибираються по порядку назви публікацій.

"Адаптація назви до пошуку" – це блок правил, який включає в себе обмеження назви, якщо вона є занадто довгою, додавання позначок для точного пошуку та домену пошуку, кодування. Наприклад, назва публікації "Аналіз кореляційної залежності між типом гідратації катіонів другої групи головної підгрупи та одно-, дво- і тризарядних аніонів та утворення кристалогідратів" буде логічно скорочена та добавлений домен і запит матиме вигляд: "Аналіз кореляційної залежності між типом гідратації катіонів другої групи головної підгрупи та одно-, дво- і" site:ena.lp.edu.ua". Досліджено, що 115 символів у запиті – це достатньо для точного знаходження публікації і для того, щоб не робити нагромадженим запит.

"Запит до Google Scholar" – передача на пошук спеціалізованій системі.

Наступним кроком є отримання результатів пошуку – "Отримання HTML-сторінки". Тоді перевіряється знайдені результати за Handle, який є ідентифікатором публікації і наявний в БД. Якщо Handle знайдено на HTML-сторінці, то це свідчить про правильність результатів пошуку на запит. Handle отримуємо з посилання (URL) публікації – цей номер є унікальний, який також використовується інформаційною системою для ідентифікації публікації. Наприклад існує URL: http://hdl.handle.net/ntb/172 з якого автоматизовано вибираємо "172". Ми відкидаємо пошук за авторами публікацій з декількох причин:

1. Google Scholar висвітлює лише першого із авторів і не відомо, який із авторів буде подаватися першим. У зв'язку з цим перевірити точність результатів до запиту стає неможливим.

2. Автор часто має різні варіанти написання, що при точному пошуку не будуть відображені адекватні результати.

3. Запит стає занадто нагромадженим, що перешкоджає пошуку.

Далі відбувається "Копіювання необхідного фрагменту", де відкидаються всі зайві дані, які були подані спеціалізованою системою як результати пошуку. Продовжується уточнення за URL, який наявний в

БД та гіперпосиланням фрагменту – це контрольний остаточний крок перевірення точності результатів пошуку.

"Наявність цитування" – це крок, де здійснюється пошук індексу цитованості (Сі) публікації, яка наявна в електронному архіві. При наявності цитування Сі := к, де к > 0 в іншому випадку Сі := 0. Знайдені значення записуються до БД. Якщо Сі = 0 – це означає, що публікація проіндексована спеціалізованою системою, але не має цитувань, відповідно, якщо Сі = к, то публікація має статті, які посилаються на неї. Якщо в БД за атрибутом "Cited" шуканої публікації існує пусте значення – це свідчить, що вона ще не проіндексована спеціалізованою системою. Записується дата пошуку.

Обсяг електронних наукових документів архіву Національного університету "Львівська політехніка" щодня поповняється новими документами. Його науковий потенціал можна аналізувати за допомогою індексів цитованості. Враховуючи регулярність досліджень та обсяги архіву постає необхідність в автоматизації пошуку індексів цитованості в спеціалізованій системі Google Scholar. Для цього використовуємо алгоритм отримання цитованості.

Такий підхід дозволяє в повній мірі аналізувати електронний архів:

– наявність цитувань наукових документів архіву;

– визначити кількість проіндексованих документів;

– швидкість індексації документів архіву роботами;

– загальну суму цитувань документів архіву та окремих видань;

– порівняння зміни суми цитованості в часі, для отримання динаміки цитованості архіву;

– побудувати рейтинг цитування публікацій;

– інші порівняння за галузями, виданнями, авторами, часовими рамками.

Висновки. В результаті аналізу наукової літератури були виділені метрики роботи алгоритму для пошуку індексів цитованості. Запропоновано здійснювати аналіз українських наукових робіт на індекс цитування, а також способи підвищення пошуку з використанням URL, handle та наявних функцій в спеціалізованій системі. Побудований алгоритм пошуку та отримання від спеціалізованої системи Google Scholar індексу цитування. Визначені етапи отримання індексу цитованості наукових електронних документів та їх функціональний опис, що дозволяє реалізувати модуль для автоматизованого отримання індексів цитованості відкритих електронних архівів. Розбиття мети, аналізу наукових електронних документів на етапи, розкриває потенційні проблеми роботи алгоритма, можливість передбачити результати аналізу, вдосконалювати методику отримання наукових показників з

спеціалізованих систем, виділити правила пошуку, обмеження та налаштування, щоб підвищити результати пошуку. Виділені напрямки аналізу електронного архіву, використовуючи запропоновану інформаційну систему. На практиці перевірено ефективність роботи алгоритму для отримання показників цитованості електронного наукового архіву Національного університету "Львівська політехніка".

Список літератури: 1. *Тарасов Д.О.* Технологічні особливості опрацювання документів у електронній формі у бібліотеках / *Д.О. Тарасов* // Інформаційні системи та мережі. – Львів: Видавництво Національного університету "Львівська політехніка", 2008. – С. 229-232. **2.** *John N.* Digital Repositories: Not Quite at Your Fingertips / *Nanchy John.* – Libri.: Germany, 2005. – P. 181-197. **3.** Using Citation Indexes [Електронний ресурс]. – Режим доступу: http://www.lib.utexas.edu/chem/info/citations.html **4.** *Adler R.* Citation Statistics / *Robert Adler, John Ewing (Chair), Peter Taylor* // A repport from the IMU in cooperation with the ICIAM and IMS, 2008. – 26 p. [Електронний ресурс]. – Режим доступу: http://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf **5.** *Noruzi A.* Google Scholar: The New Generation of Citation Indexes / *Alireza Noruzi.* – Libri.: Germany, 2005. – P. 170-180. **6.** *Archambault, E.* The Use of Bibliometrics in Social Sciences and Humanities / *Archambault E., Gagné E.V. //* Montreal: Social Sciences and Humanities Research Council of Canada (SSHRCC) . – Canada, 2004. – 72 p. **7.** Citation Metrics [Електронний ресурс]. – Режим доступу: http://library.amnh.org/research-tools/tips-tutorials/citation-metrics **8.** How to Search Google Scholar [Електронний ресурс]. – Режим доступу: http://www.mslib.huji.ac.il/en/faq/google-scholar.html

Статью представил д.т.н., доц. Національного університету "Львівська політехніка" Пелещин А.М.

УДК 51.001.57

Алгоритм получения данных для построения индекса цитируемости документов открытых электронных архивов / Тарасов Д.О., Гарасим О.Р. // Вестник НТУ "ХПИ". Серия: Информатика и моделирование. – Харьков: НТУ "ХПИ". – 2012. – № 38. – С. 190 – 195.

Обоснована необходимость создания алгоритма для получения цитируемости научных документов открытых электронных архивов. Построен алгоритм получения индекса цитируемости в специализированной системы Google Scholar. Описаны основные шаги работы алгоритма. Ил.: 1. Библиогр.: 8 назв.

Ключові слова: Google Scholar, индекс цитируемости, алгоритм получения индекса цитируемости.

UDC 51.001.57

The algorithm to obtain data for constructing citation index documents of open electronic archives / Tarasov D.O., Garasym O.R. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2012. – P. 190 – 195

The necessity of creating an algorithm to obtain the citations of scientific documents open electronic archives. An algorithm to obtain the citation index of a specialized system Google Scholar is constructed. Describes the main steps operation of the algorithm. Figs.: 1. Refs.: 8 titles.

Keywords: Google Scholar, citation index, algorithm for citation obtaining.

Надійшла до редакції 27.06.2012