УДК 004.62

МЕТОДИ ПОШУКУ ДАНИХ НА ПОРТАЛІ НАЦІОНАЛЬНОГО АНТАРКТИЧНОГО НАУКОВОГО ЦЕНТРУ УКРАЇНИ

Новогрудська Р.Л., Юшко Н.А.
Інститут телекомунікаційних систем КПІ ім. Ігоря Сікорського, Україна
E-mail: rinan@ukr.net, natalia.yushko@outlook.com

DATA SEARCH METHODS ON THE NATIONAL ANTARCTIC SCIENTIFIC CENTER OF UKRAINE

The relevance of the work is to resolve the problem of organizing an effective search for a large number of poorly structured and heterogeneous information in the National Antarctic Data Center (NADC). The development and implementation of the right search method can improve the efficiency and speed of the system. On this basis, the aim of the work is to increase the efficiency and speed of the search for Antarctic research results stored in the National Antarctic Data Center. When selecting the method of searching, it is necessary to take into account the specificity of the data formats of the Antarctic researches located in the NADC, as well as their structuring and type.

Щорічно науковці ДУ НАНЦ проводять наукові експедиції на станцію Академік Вернадський. З кожної поїздки науковці антарктичного центру привозять 1000 – 2000 Гб необроблених даних. В свою чергу оброблені дані займають ще 200 – 300 Мб. Крім щорічних поїздок на рік, існують також і зимові поїздки на три місяці. В результаті таких зимівок дослідниками також збирається певний об’єм інформації [1]. Задача ускладняється тим, що співробітники працюють з даними різної форми (текст, медіа-, відеофайли, файли інших розширень). Тому, для того, щоб оптимально вирішити задачу зберігання інформації, необхідно підібрати відповідний метод пошуку інформації для підвищення ефективності роботи з порталом НЦАД. З цією метою необхідно розглянути поняття пошук інформації.

Інформаційний пошук (ІП) – процес пошуку неструктурованої документальної інформації, що задовольняє інформаційні потреби, та наука про пошук неструктурованої документальної інформації. Особливо це відноситься до пошуку інформації в документах, пошук самих документів, здобуття метаданих з документів, пошуку тексту, зображень, відео та звуку у локальних реляційних базах даних, у гіпертекстових базах даних таких, як Інтернет та локальні Інтранет, на веб-ресурсах та у електронних каталогах даних[2]. Пошук інформації являє собою процес виявлення в деякій множині документів (текстів) всіх тих, які присвячені зазначеної теми (предмету), задовольняють заздалегідь визначеним умовам пошуку (запиту) або містять необхідні (відповідні інформаційної потреби) факти, відомості, дані. Процес пошуку включає послідовність операцій, спрямованих на збір, обробку та надання інформації. У загальному випадку пошук інформації складається з чотирьох етапів: визначення (уточнення) інформаційної потреби і формулювання інформаційного запиту; визначення сукупності можливих власників інформаційних масивів (джерел); вилучення інформації з виявлених інформаційних масивів; ознайомлення з отриманою інформацією і оцінка результатів пошуку. На даний момент часу розрізняють такі типи

інформаційного пошуку як повнотекстовий пошук, пошук по метаданим, пошук по зображенням.

Повнотекстовий пошук - автоматизований пошук документів, при якому пошук ведеться не по іменах документів, а по їх вмісту, всьому документу або по його частині [3]. Режим повнотекстового пошуку дозволяє ввести в пошуковий рядок одне слово, а пошук тексту буде здійснюватися по всіх його словоформам. Задля того, щоб забезпечити найбільшу продуктивність системи, необхідно ввести поняття індекс пошуку. Повнотекстовий індекс - словник, в якому перераховані всі слова наявні у даній системі чи документі, та вказано, в яких місцях вони зустрічаються. Повнотекстовий запит повертає всі документи, які містять як мінімум один збіг (відоме також як потрапляння).

Пошук по метаданим - це пошук за деякими атрибутам документа, підтримуваним системою: стандартними - такими як назва документа, дата створення, розмір, автор, або ж атрибутами що визначає безпосередньо користувач - наприклад, категорія документу, номер поїздки (зимівки), відповідальна особа і т.д. [4]

Пошук по зображеннях - пошук по змісту зображення. Пошукова система розпізнає вміст фотографії (завантажена користувачем або додана URL-адреса зображення). [5]

Так як пошук необхідно здійснити по Національному Антарктичному Центрі Даних, який в свою чергу побудований на основі платформи Microsoft SharePoint 2016, розглянемо метод пошуку який належить до SharePoint. Пошук в SharePoint можна віднести до інформаційного пошуку, що виконує повнотекстовий пошук, включаючи пошук по метаданим та по різнотипій інформації. Даний метод пошуку використовує індексування елементів (використання пошукового індексу файлів) для підвищення ефективності пошуку. Пошук в SharePoint реорганізовано в єдину корпоративну пошукову платформу. Архітектура пошуку включає наступні області: обхід та обробка контенту; індексування елементів; обробка запитів пошуку; адміністрування пошуку; аналітика пошукових запитів. При початку роботи пошукового алгоритму виконується обхід контенту, компонент обходу збирає властивості для обходу та метадані з обхідних елементів і передає їх у компоненти обробки контенту. База даних обходу містить інформацію про оброблені елементи, наприклад, останній час обходу контенту, ідентифікатор останнього обходу і тип поновлення під час останнього обходу. Компонент обробки контенту обходить джерела контенту, щоб зібрати властивості і метадані з обійдених компонентів, та відправляє отриману інформацію в компонент індексування. Компонент індексування отримує від кодека контенту оброблені елементи і записує їх в індекс пошуку. Крім того, цей компонент обробляє вхідні запити, отримує інформацію від пошукових індексів і відправляє набір результатів назад компоненту обробки контенту. Компонент обробки запитів аналізує і обробляє пошукові запити і результати. Після чого оброблений запит відправляється в компонент індексування, який повертає набір результатів пошуку для даного запиту [6]. Детальніше алгоритм роботи даного методу пошуку зображений на рисунку 1.

Веб-частина пошуку контенту це веб-частина, що відображає динамічний контент, який раніше був обійдений і доданий в індекс пошуку. Кожен

екземпляр веб-частини пов'язаний з індексом пошуку і відображає результати для даного конкретного запиту. Коли користувачі переглядають сторінку, яка містить веб-частину пошуку контенту, автоматично видається запит, і від індексу пошуку повертаються результати пошуку.

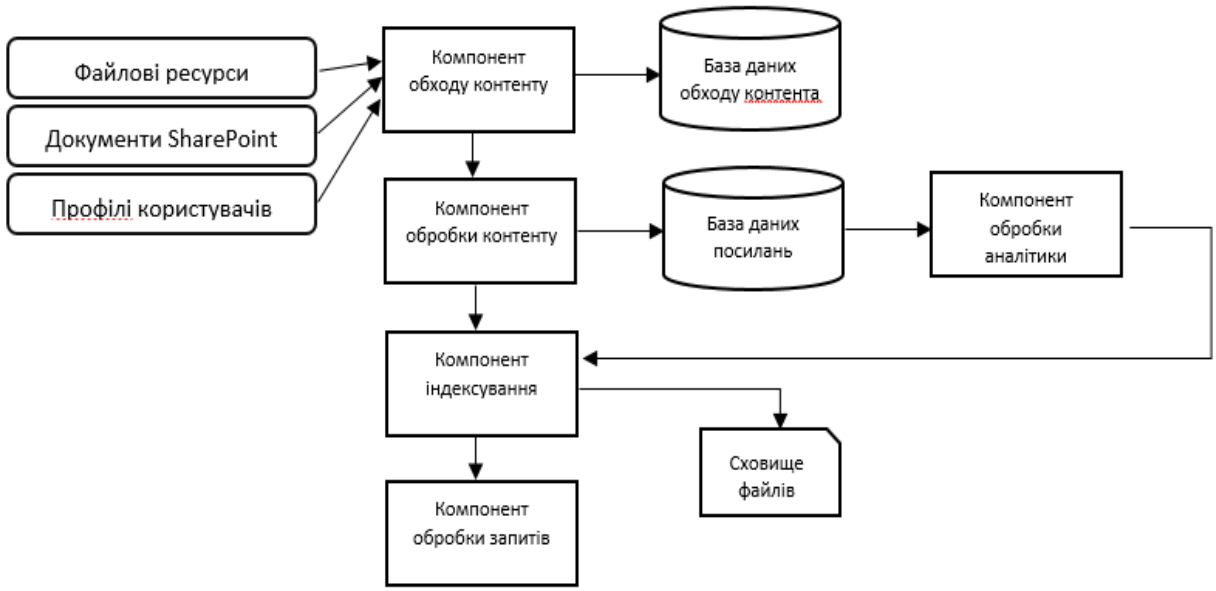

Рис. 1. Алгоритм пошуку в SharePoint.

Таким чином, проаналізувавши даний алгоритм пошуку можна зробити висновки, що він є найоптимальнішим для реалізації на порталі Національного Антарктичного Центру Даних. Адже за допомогою нього можна реалізувати ефективний пошук не лише по всьому вмісту всіх стандартних типів файлів, а й виконати пошук по всім стандартним та користувацьким метаданим. Заданий алгоритм пошуку забезпечує швидке оброблення великої кількості інформації за рахунок застосування індексування, та забезпечує подальшу аналітику пошукових запитів за рахунок наявності компонента обробки запитів. Даний алгоритм вже присутній на платформі SharePoint 2016, що спрощує задачу його застосування для Національного Антарктичного Центру Даних.

Література

1. Глоба Л.С., Мороз І. В., Новогрудская Р.Л., Мочалкина К.С., Кузін І.О.Создание единого информационного пространства данных антарктических исследований, Український Антарктичний Журнал, №10-11, 2011, с. 343-351.
2. Ланде Д. В., Снарский А. А., Безсуднов И. В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. — М.: Либроком (Editorial URSS), 2009. — 264 с.
3. Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval. — Cambridge University Press, 2007.
4. [Електронний ресурс] Інформаційний пошук даних на веб ресурсах та електронних каталогах http://www.viaduk.net/viaduk/notes_client.nsf/f4b82fbb75e942a6852566ac0037f284/25fa84a86061aaafc32570c800598252?OpenDocument Дата доступу: 19.02.18.
5. Shapiro Linda. Computer Vision. — Upper Saddle River, NJ: Prentice Hall, 2001.
6. [Електронний ресурс] Пошук в SharePoint https://docs.microsoft.com/uk-ua/sharepoint/dev/general-development/search-in-sharepoint#bk_crawl, Дата доступу 20.03.18.