Шамота В.О., магістрант, група ПІ-48М,
Науковий керівник – Сугоняк І.І., к.т.н., доцент,
Житомирський державний технологічний університет

WEB-ОРІЄНТОВАНА СИСТЕМА ПОШУКУ ТА АНАЛІЗУ ТЕМАТИЧНОГО КОНТЕНТУ

В сучасному розумінні, пошукова система – це дуже складний програмний комплекс, алгоритми роботи якого тримаються в суворій таємниці його розробниками. За допомогою цих самих пошукових алгоритмів пошукові системи збирають, а потім індексують всю інформацію про веб-сторінки. Для цього зібрану інформацію заносять в спеціальну базу даних, проводять її структуризацію та розташовують у необхідному порядку. Використання ПС споживачем інтернету через спеціальні програми – браузери, – це ніщо інше, як звернення його до цієї бази даних.

Основне завдання пошукової системи – швидко сформувати сторінку з правильними відповідями на запит користувача. На перший погляд, це нескладне завдання, але якщо врахувати кількість користувачів, кілька сотень способів введення пошукових запитів і десятки мільйонів сайтів – задача суттєво ускладняється.

Пошукові системи можуть працювати повністю за допомогою роботів або людей, а також представляти із себе гібридні системи.

В архітектуру пошукової системи найчастіше входять: робот, індексатор та пошуковик. Перший відповідає за збір інформації з різних документів (сторінок сайтів), індексатор – за швидкий пошук інформації, а останній – формує сторінку з результатами пошуку.

Системи намагаються дати не тільки посилання на корисні для користувача документи, а й частково сформувати відповідь з різних джерел відразу на сторінці видачі.

Основними характеристики якісного пошуку є: повнота – чим більше проаналізованих документів, тим повніше пошук; точність – користувач не захоче шукати голку в стозі сіна, результат має бути релевантним, актуальність – особливо важливо при пошуку новин; швидкість пошуку – ніхто не буде чекати, поки система згенерує сторінку з відповідями; наочність – зручне представлення інформації.

В даній роботі розглядається розробка WEB-орієнтованої системи для пошуку та аналізу тематичного контенту в мережі Інтернет.

Наукова новизна представленої роботи полягає у постійній необхідності пошуку користувачами інформації в мережі Інтернет, а також, розробці нових алгоритмів для покращення можливостей пошуку, збільшенню його точності, отриманню максимальної кількості результатів, задовольняючи пошуковий запит та, найголовніше, актуальності отриманих результатів.

Система буде складатися із трьох компонентів:

1) WWW-сервер. Відповідає за взаємодію пошукової системи з користувачем. Надає зручний і наочний інтерфейс для завдання запитів.
2) Пошукова машина. Обробляє базу індексів відповідно до отриманого запиту.
3) Пошуковий робот. Комп'ютер, оснащений спеціальною програмою, яка безперервно переглядає весь Інтернет, індексуючи всі Web-сторінки, і оновлюючи базу індексів.

Індексатор повинен виконувати наступні дії: пошук файлів заданого типу на диску; виділення заголовку із тегів <TITLE> та </TITLE>; виділення інформації із документів, фільтруючи скрипти (теги <SCRIPT> та </SCRIPT>), а також таблиці стилів (теги <STYLE> та </STYLE>); збереження інформації про документи в спеціальному файлі, присвоєння документу унікального номера; збереження інформації про кожне слово, і номер документа в індексі.

Механізм пошуку за індексом повинен виконувати наступні дії: шукати в індексному файлі кожне слово з пошукового запиту; сортувати результати пошуку за релевантністю і відповідностю, тобто, в першу чергу виводяться результати, що строго відповідають запиту (якщо кожне слово запиту існує в знайденому документі), також, результати сортуються за кількістю слів запиту, зустрінутих в знайденому документі; форматувати результати пошуку і розбивати їх на сторінки.