УДК 004.051
Шаповалова А.– ст. гр. СІм-52
Тернопільський національний технічний університет імені Івана Пулюя

ДОСЛІДЖЕННЯ МЕТОДІВ І АЛГОРИТМІВ ПОШУКУ В WEB-ПОШУКОВИХ СИСТЕМАХ

Науковий керівник: к.т.н., доцент Шингера Н.Я.

Shapovalova A.
TernopilIvanPul'ujNationalTechnicalUniversity

RESEARCH METHODS AND ALGORITHMS FOR SEARCHING THE WEB-SEARCH ENGINES

Supervisor: Shynhera N.Ya.

Ключові слова: пошукова система, ранжування, алгоритм, метод
Keywords: searchengine,ranking,algorithm, method

На ранніх стадіях еволюції в алгоритмі web-пошукових систем враховувалася незначна кількість факторів, що впливала на ранжування у видачі результатів пошуку, тому, знаючи базові принципи роботи пошукових систем, можна було досить легко маніпулювати результатами, що й робили багато компаній, які займалися просуванням Інтернет ресурсів. Для того щоб підтримувати якість пошуку, найважливішого аспекту, пошукові системи були змушені ускладнювати свої алгоритми - кількість чинників, що враховуються зросли в сотні, і навіть тисячі разів, стали все частіше з'являтися різні алгоритми і фільтри. Згодом, число таких чинників зростало в геометричній прогресії, постійно збільшувався рівень конкуренції між пошуковими системами, виводячи на пошуковий ринок тільки тих, хто міг боротися і надавати користувачам Інтернету релевантні результати незалежно від вміння seo-фахівців просувати сайти.

Ресурси Інтернету перетворилися в незамінний інструмент для повсякденної роботи людей багатьох професій. Швидке зростання інформації в мережі зробили його океаном найрізноманітніших даних, важливість яких зростає пропорційно їх обсягу. За оцінкою спеціалістів об'єм інформації, що передається по каналах всесвітньої павутини, подвоюється кожні півроку. Щодня в мережі з'являються сотні тисяч нових документів, і очевидно, що без систем пошуку вони залишалися переважно незатребуваними,або не знаходилися взагалі. Тому, виникла необхідність створення таких засобів, які дозволили б легко орієнтуватися в інформаційних ресурсах глобальних мереж, швидко і надійно знаходити потрібні дані.

Пошукова система - це потужний апаратно-програмний комплекс, що призначений для здійснення пошуку ресурсів в Інтернеті. Основну мету, яку переслідують їх розробники, є індексування сторінок і документів в глобальній мережі для подальшого надходженнята видачі відповідно до запитів користувачів. Для того, щоб надана інформація була актуальною та якісною, розробники невпинно вдосконалюють формули та методи ранжування.

Завдання пошукової системи полягає в тому, щоб при видачі результатів пошуку забезпечити максимальний збіг слів в пошуковому запиті зі словами, знайденими на тій чи іншій веб-сторінці або в тексті посилань, вказуючих на неї.

Ранжування - це процес, при якому пошукові системи сортують сайти на сторінці результатів пошукув певному порядку за ступенем їх важливості, значущості.

Зазвичай пошуковики тримають в таємниці свої критерії ранжування, проте найбільш загальновідомими характеристиками є наступні:

- кількість слів/словосполучень, представлених в запиті користувача, на шуканій веб-сторінці;
- розташування слів/словосполучень запиту в знайденому документі (заголовки, текст, виділений текст і т.д.);
- пряма відповідність формам слів в запиті (відмінок, число, частина мови і т.д.);
- відстань між словами, зазначеними в запиті, і словами на знайденій веб-сторінці;
- контрольна вага веб-сторінки, тобто наскільки часто на дану сторінку посилаються інші веб-ресурси за вказаним запитом.

Метою роботи є дослідження існуючих алгоритмів та методів пошуку в сучасних пошукових системах за показниками якості для виявлення основних факторів, що впливають на ранжування сайтів в результатах пошуку.

Досягнення поставленої мети передбачає розв’язання наступних завдань:

- провести аналіз сучасних пошукових систем, використовуваних в українському і світовому сегменті мережі Інтернет;
- розробити систему факторів, що беруть участь в формулах ранжування пошукових систем;
- здійснити огляд алгоритмів пошукових систем;
- розглянути теоретичні підходи до обгрунтування проблеми пошукової оптимізації та ранжування web-сайтів;
- з'ясувати сутність пошукової оптимізації та ранжирування як предметів дослідження, охарактеризувати чинники ранжування;
- дослідити специфіку роботи сучасних пошукових машин;
- надати рекомендації щодо розробки програмного модуля на основі власного алгоритму пошукової системи.

Практична цінність роботи полягає в можливості використання отриманих науково-технічних результатів при експлуатації, дослідженні, що вимагають відносного порівняння альтернативних алгоритмів і методів.

В якості досліджуваних пошукових систем, для яких проводилася оптимізація, були обрані найбільш популярні та прогресивні Яндекс і Google.

Сучасні алгоритми Google використовують понад 200 різних сигналів або "ключів", щоб зрозуміти, що саме шукає користувач. Враховується такі параметри, як наявність слів на сторінках сайтів, актуальність інформації, місце розташування користувача і показник PageRank інтернет ресурсів і т.д.

PageRank - один з алгоритмів посилання ранжування. Алгоритм застосовується до колекції документів, пов'язаних гіперпосиланнями (таких, як веб-сторінки з всесвітньої павутини), і призначає кожному з них якесь чисельне значення, що вимірює його «важливість» або «авторитетність» серед інших документів. Крім того, «вага» сторінки A визначається вагою посилання, переданої сторінкою B. Таким чином, PageRank - це метод обчислення ваги сторінки шляхом підрахунку важливості посилань на неї. Google використовує показник PageRank знайдених за запитом сторінок, щоб визначити порядок видачі цих сторінок користувачу в результатах пошуку.

Для вирішення поставлених завдань використані методи теорії множин, теорії систем масового обслуговування, порівняльного аналізу, об'єктно-орієнтованого аналізу, розрахована спрощена формула для визначення релевантності сторінки сайту, за допомогою схем, графіків, діаграм і спостережень.