УДК 004.891

ВДОСКОНАЛЕННЯ АЛГОРИТМУ РАНЖУВАННЯ ТА ІНДЕКСАЦІЇ САЙТІВ

Мясіщев О.А., Судома І.В.
Хмельницький національний університет

У статті запропоновано модифікацію алгоритму ранжування Google – PageRank. Модифікація дозволить давати більш відповідну вагу популярності сайтам. При пошуку в мережі Інтернет, користувач зможе отримати відсортований, проіндексований список результатів із сайтами та необхідну йому інформацію. Суть та функціональність модифікованого алгоритму полягає в тому, що буде отримана нова формула розрахунку ваги сайту – SocPageRank. Вона отримає додатковий коефіцієнт та змінну, на основі яких буде відбуватись порівняльна характеристика реальної оцінки сайту із популярністю цього ж сайту на форумах та соціальних мережах.
Ключові слова: інформація, алгоритм, пошук, важливість, PageRank, SocPageRank, сайт, пошукова система.

Постановка проблеми. Для виявлення недоліків видачі результатів запиту на сторінку користувача, потрібно розглянути як саме побудований PageRank.

Порядок ранжування сайтів в Google працює наступним чином [1]:

1) знайти всі сайти які відповідають ключовим словам пошуку;

2) відранжувати відповідно до «сторінкових факторів», таких як ключові слова;

3) взяти до уваги текст посилань на сайти;

4) відкорегувати результати даними PageRank.

Декілька факторів про PageRank. PageRank – це число, що характеризує виключно здатність голосування всіх вхідних посилань на сторінку і те як вони рекомендують цю сторінку [2]. Кожна унікальна сторінка сайту, проіндексована Google, має вагу PageRank. Люди часто помиляються, думаючи про вагу сайту, який насправді є вагою головної сторінки цього сайту. Внутрішні посилання сайту враховуються при розрахунку ваги PageRank для інших сторінок сайту. PageRank незалежний, він не бере до уваги текст посилань і т.д.

Проблема отримання потрібних та корисних джерел була актуальною протягом усього розвитку інформаційних технологій, отже має місце на існування модифікований алгоритм пошуку інформації SocPageRank.

Аналіз останніх досліджень і публікацій. Коли Google був тільки дослідницьким проектом, Брін і Пейдж написали статтю яка докладно описує формулу, що визначає вагу PageRank для сторінки. Хоча вони, можливо, вже не використовують в точності цю формулу, вона є досить коректною для сьогоднішніх цілей. Ось вона [2]:

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)}\right), \quad (1)$$

де $PR(A)$ – вага сторінки А (те, що ми шукаємо);

D – коефіцієнт затухання, який зазвичай встановлюють рівним 0,85;

$PR(T_1)$ – вага *PadeRank* сторінки, що посилається на сторінку А;

$C(T_1)$ – кількість посилань із цієї сторінки;

$\frac{PR(T_n)}{C(T_n)}$ – для кожної сторінки, яка вказує на сторінку А.

Одним з перших показників посилання ранжування, заснованим на передачу так званої ваги посилання, став саме цей алгоритм. Згодом цей алгоритм удосконалювався творцями кожної з пошукових систем, ускладнювався і все менше впливав на загальну релевантність документа. Однак в усі посилальні алгоритми пошукових систем закладена ідея PageRank, створена в 1996 році Сергієм Бріном і Ларрі Пейджем, вдосконалена і ускладнена.

Виділення не вирішених раніше частин загальної проблеми. Одна людина може знайти необхідну їй інформацію за декілька секунд і проаналізувати видані результати ввівши в адресну стрічку адресу сайту, інша зробить це ввівши конкретно сформований та цілеспрямований запит у пошуковій системі, а є і такі користувачі які не мають необхідних здібностей та аналітичного мислення для того щоб знайти потрібний їм ресурс.

Саме в таких користувачів дуже часто виникають проблеми із знаходженням необхідного джерела та інформації відповідної до введеного запиту у пошукове поле. Для боротьби із такими проблемами пошукові системи звичайно впроваджують такі міри як: поправка введеного тексту у поле запиту, інтелектуальний пошук, випадаючі допоміжні слова та популярні запити в процесі формування запиту. Але, все ж, це не забезпечує більшої ефективності в процесі пошуку. Саме вирішення проблеми видачі адекватних та корисних результатів пошуку буде розглянуто у цій статті.

Мета статті. Метою статті є дослідження та вдосконалення функціонування алгоритму ранжування PageRank.

Виклад основного матеріалу. Для отримання необхідних корисних результатів пошуку в мережі та отримання реальних оцінок сайту, PageRank буде модифіковано ще одним коефіцієнтом який відповідатиме за популярність сайту у соціальних мережах та його рекомендування на форумах.

Тобто у формулу ранжування PageRank буде додано змінну s – коефіцієнт який буде рівний 0,15, буде множитись на кількість посилань з форумів та соціальних мереж – $T(C_s)$.

Повернемося до математичного рейтингу сайту, та розглянемо як у цій простій мережі функціонує алгоритм ранжування.

Веб-сайт С має більш високий рейтинг, ніж сайт Е, хоча є менше посилань на С, ніж на Е, але одна з посилань на С виходить з важливіших сайтів і, отже, має більш високе значення. Якщо умовно вважати, що веб-користувач, який знаходиться на випадковому сайті, має 85% ймо-

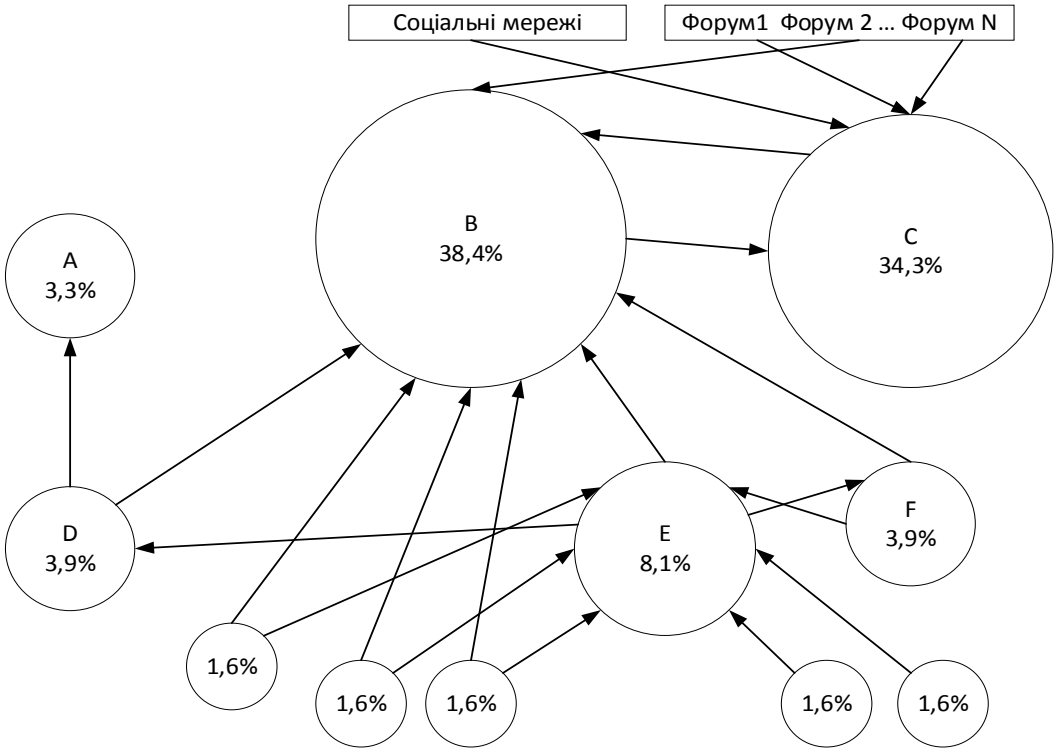

Рис. 1. Модифікована модель математичного рейтингу сайту

Джерело: розроблено автором

вірність вибору випадкового посилання на поточному сайті, і 15% переходу на будь-який інший сайт, то ймовірності переходу на сторінку E з інших посилань дорівнює 8,1% часу. (15% ймовірності переходу до довільного сайту відповідає коефіцієнту загасання 85%). Без загасання вебкористувачі в кінцевому підсумку потрапляють на сайти A, B або C, і всі інші сайти будуть мати PageRank, рівний нулю. При наявності загасання сайт A ефективно пов'язує майже всі посилання на сайти в цій мережі, навіть якщо він не має своїх власних вихідних посилань.

Отже модифікований алгоритм ранжування SocPageRank, набуде вигляду як зображено на рисунку 1.

Як видно з рисунку 1, після впровадження коефіцієнта s реальна оцінка сайту B, дещо впаде, оскільки він не настільки популярний на форумах та соціальних мережах як сайт C. До оцінки сайту C доплюсується 0,45:

$s \times T(C_s) = 3 \times 0{,}15 = 0{,}45$. Відповідно до загальної оцінки сайту B доплюсується: $s \times T(C_s) = 1 \times 0{,}15 = 0{,}15$.

В такому випадку сайти B та C стануть рівними за оцінками, а отже, відповідно до модифікованого алгоритму, перевага у списку видачі результатів пошуку піде на сторону сайту C.

За такого розкладу подій, коли на форумах сайт C обговорюють та рекомендують частіше та більше ніж сайт B. Проблема може бути в тому, що сайт B дійсно не є настільки популярним завдяки своїй корисності та актуальності.

Висновки і пропозиції. Всі пошукові системи борються з просуванням сайтів чорними методами оптимізації. Сайти, помічені у використанні чорних способів оптимізації, досить часто навіть виключаються з індексу на якийсь термін, або назавжди.

Пошукові системи існують для спрощення роботи користувача з інформацією з Всесвітньої павутини.

Проблема отримання потрібних та корисних джерел була актуальною протягом усього розвитку інформаційних технологій, отже має місце на існування модифікований алгоритм пошуку інформації SocPageRank.

Список літератури:

1. Как работает Google поиск [Электронный ресурс]. – Режим доступа: https://habr.com/company/ua-hosting/blog/277819/.
2. Растолкованный PageRank: или все, что вы всегда хотели знать о PageRank [Электронный ресурс]. – Режим доступа: http://digits.ru/articles/promotion/pagerank.html.

Мясищев А.А., Судома И.В.
Хмельницкий национальный университет

УСОВЕРШЕНСТВОВАНИЕ АЛГОРИТМА РАНЖИРОВАНИЯ И ИНДЕКСАЦИИ САЙТОВ

Аннотация
В статье предложен модификацию алгоритма ранжирования Google – PageRank. Модификация позволит давать более соответствующий вес по пулярности сайтам. При поиске в сети Интернет, пользователь сможет получить отсортированный, проиндексированный список результатов с сайтами и необходимую ему информацию. Суть и функциональность модифицированного алгоритма заключается в том, что будет получена новая формула расчета веса сайта – SocPageRank. Она получит дополнительный коэффициент и переменную, на основе которых будет происходить по-сравнительный характеристика реальной оценки сайта с популярностью этого же сайта на форумах и социальных сетях.
Ключевые слова: информация, алгоритм, поиск, важность, PageRank, SocPageRank, сайт, поисковая система.

Myasishchev O.A., Sudoma I.V.
Khmelnitskiy National University

IMPROVING THE ALGORITHM OF RANKING AND SITE INDICATIONS

Summary
The article proposes a modification of Google's ranking algorithm – PageRank. The modification will allow to give a more appropriate weight to the sites of sites. When searching the Internet, the user will be able to receive a sorted, indexed list of results with the sites and the information it needs. The essence and functionality of the modified algorithm is that we will get a new formula for calculating the weight of the site – SocPageRank. It will receive an additional coefficient and a variable, which will be based on the level characteristic of the real assessment of the site with the popularity of the same site on forums and social networks.
Keywords: information, algorithm, search, importance, PageRank, SocPageRank, site, search engine.