УДК 004.738.5

ПЕРСПЕКТИВИ ВДОСКОНАЛЕННЯ ІНФОРМАЦІЙНОГО ПОШУКУ

Терещенко В.В.
Кременчуцький національний університет
імені Михайла Остроградського
Терещенко В.Л.
Філія «Кременчуцький інститут
ВНЗ «Університет імені Альфреда Нобеля»

В сучасних інформаційних технологіях мережі Internet та пошукових машинах є потреба у вдосконаленні пошукових алгоритмів. У рамках підвищення ефективності пошукової видачі запропоновано: використання моделі векторного простору (VSM); метод винятковостей (англ. single method). Для підвищення достовірності оцінки документу наголошено на доцільності використання методу SeoRank. Результати дослідження будуть корисні при вдосконаленні пошукових алгоритмів.
Ключові слова: інформаційний пошук, пошукова система, пошукова видача, релевантність, пошуковий алгоритм, веб-сторінка.

Постановка проблеми. В умовах науково-технічного прогресу і розвитку Інтернет-технологій відбувається надзвичайне зростання обсягів доступної інформації, яка може бути використаною при вирішенні важливих завдань в ході науково-дослідної діяльності, для підтримки прийняття рішень в науково-технічній, соціальній та інших сферах [1]. Ефективний аналіз цієї інформації та її застосування при прийнятті стратегічних рішень дає перевагу в розвитку не лише економіки, а й науки та технологій.

Так як вимоги до швидкості пошуку, актуальності інформації з кожним днем зростають, то збільшуються і вимоги до методів та алгоритмів пошуку і подання інформації. Процес пошуку та відображення інформації в Інтернеті має ряд особливостей, головними з яких є величезна кількість веб-ресурсів, необхідність врахування семантичних особливостей інформації, вплив великої кількості факторів при пошуку, необхідність врахування особливостей гіпертекстової розмітки та метаінформації [1]. На сьогоднішній день існує значна кількість методів та алгоритмів інформаційного пошуку, проте неперервний розвиток цієї галузі та зростання обсягів даних вимагає постійного покращення існуючих методів та розробку якісно нових підходів. Тож проблема вдосконалення алгоритмів інформаційного пошуку є актуальною.

Аналіз останніх досліджень і публікацій. В рамках цієї проблеми працювали багато дослідників: Ашманов І.С. [1], Колісниченко Д.М. [2], Крохіна О.І. [3], Маннінг К.Д. [4] та ін. Так, наприклад в книзі І.С. Ашманова [1] узагальнено досвід відомих фахівців, SEO-професіоналів; особливої уваги заслуговує аналіз принципів роботи пошукових систем. Д.М. Колісниченком [2] докладно описані алгоритми роботи і методи використання найбільш популярних сьогодні пошукових машин Інтернету – Google, Яндекс і Рамблер. Окрім цього, автором розглядаються способи розробки власних Google-додатків: особистих пошукових машин, створених на базі інструментальних засобів Google. Не зважаючи на те, що робота О.І. Крохіної [3] орієнтована на SEO-копірайтерів, інтернет-маркетологів, фахівців з пошукової оптимізації, веб-майстрів і власників сайтів, у ній розглядаються загальні приншипи роботи пошукових алгоритмів. Саме спираючись на них вона пояснює як написати текст для сайту, який однаково добре буде сприйматися користувачами і забезпечить високі позиції у видачі пошукових систем. Незважаючи на те що підручник К.Д. Маннінга [4] задуманий як вступний курс з інформаційного пошуку та написаний з точки зору інформатики; в ньому поряд з класичним пошуком розглядаються веб-пошук, принципи роботи пошукових механізмів а також класифікація та кластеризація текстів. Книга містить сучасний виклад всіх аспектів проектування та реалізації систем збору, індексування та пошуку документів, методів оцінки таких систем, а також введення в методи машинного навчання.

Очевидно, що проблема широко обговорюється науковим співтовариством. Однак, попри значну кількість публікацій дослідників, проблема вдосконалення алгоритмів інформаційного пошуку не розв'язана повністю та залишається актуальною.

Виділення не вирішених раніше частин загальної проблеми. Суть інформаційного пошуку у загальному випадку зводиться до того, що пошукова система за певними критеріями обирає з безлічі документів, що знаходяться в базі, такі, які задовольняють інформаційну потребу і відповідають інформаційному запиту (тобто є релевантними) [4]. Відповідно до загальних принципів організації інформаційного пошуку, в основі кожного пошукового алгоритму знаходиться модель реалізації, яка використовується для деталізації пошукової стратегії [2]. Таким чином, можна сказати що формально для пошукового алгоритму вона є математичним представленням, здатним відобразити будь-який релевантний об'єкт в інформаційно-пошуковій системі співвідносно з будь-якими критеріями його використання системою, з метою виконання пошукового завдання.

Проблема в тому, що відмінності між існуючими пошуковими алгоритмами породили значну різноманітність моделей [2]. Якщо модель буде досить загальною, то відповідний пошуковий алгоритм буде корисним лише для дуже поверхневої концептуалізації інформаційного пошуку. З іншого боку, якщо модель буде визначеною досить глибоко, щоб охопити всі можливі аспекти системи, то виникне проблема в складному описі принципів організації, що створюватиме труднощі

для подальшого вдосконалення алгоритму. Таким чином, доцільним буде створення вдосконаленого алгоритму реалізації інформаційного пошуку, модель якого буде однаково відповідною як критеріям загальності, так і критеріям глибинності.

Мета статті. Головною метою даної роботи є суттєве поліпшення результатів інформаційного пошуку за показниками релевантності.

Виклад основного матеріалу. Враховуючи значне накопичення об'ємів інформації в сукупності з прогресуючими темпами зростання її кількості та важливість відкриттів в області інформаційного пошуку актуальним залишається питання розробки та вдосконалення пошукових алгоритмів.

Оптимізація інформаційного пошуку з'явилася в період розвитку пошукових систем [1]. У той час пошукові системи надавали велике значення аспектам, якими власники сайтів могли легко маніпулювати: текст на сторінці, ключові слова в мета-тегах та інші внутрішні чинники. Це призвело до того, що у видачі багатьох пошукових систем перші кілька сторінок займали сайти, які були повністю присвячені рекламі [2].

Ідея автоматизованої обробки текстової інформації за допомогою електронно-обчислювальних машин виникла ще на початку XX століття. Розвиток комп'ютерної лінгвістики сприяв інтеграції методів математики (перш за все, статистики та дискретної математики) та лінгвістики для вирішення прикладних завдань аналізу текстової інформації [3].

Так, з появою Google PageRank [7] більше уваги стало надаватися зовнішнім факторам, що допомогло Google стати лідером пошуку у світовому масштабі, ускладнивши оптимізацію за допомогою лише тексту на сайті. Впродовж довгого часу PageRank був одним з найголовніших алгоритмів ранжування Google [1]. Згодом модифікований алгоритм застосовувався до колекції документів, пов'язаних гіперпосиланнями (таких, як веб-сторінки з всесвітньої павутини), і визначав кожному з них деяке чисельне значення, що вимірювало його «важливість» або «авторитетність» серед інших документів.

Чим більше існувало посилань на сторінку, тим «важливішою» вона була. Крім того, «вага» сторінки А визначалася вагою посилання, переданою сторінкою В. Таким чином, PageRank був методом обчислення ваги сторінки шляхом підрахунку важливості посилань на неї [3].

Однак, для вирішення більш складних завдань інформаційного пошуку (комп'ютерний переклад, автоматичне реферування та інші завдання аналітичної обробки текстової інформації) необхідно використовувати методи лінгвістичного аналізу текстів, які надають змогу не лише виявляти поняття, ключову лексику а і дозволять визначати різні зв'язки між ними.

Враховуючи вищезазначене, перспективним в плані загальності та глибинності при роботі з текстом буде використання моделі векторного простору (VSM) [4]. За допомогою відповідної моделі описуватиметься алгоритм інформаційного пошуку за модифікованим частотним критерієм, який окрім використання релевантності слова враховуватиме також його семантичну вагу, покращуючи тим самим якість пошукового запиту. Це надасть змогу отримувати релевантні дані навіть у тому випадку, коли більшість слів запиту не містяться у контексті (документі), незважаючи на семантичну подібність між контекстом та запитом. Vector Space Model (VSM) – це математична модель [4] представлення текстів, в якій кожному документу зіставлений вектор, що виражає його зміст. Таке уявлення дозволяє легко порівнювати слова, шукати схожі, проводити класифікацію, кластеризацію і т.д.

У загальному випадку існують два основні підходи до семантичного пошуку, та й взагалі до порівняння документів за змістом. Перший підхід заснований на ручному наділенні об'єктів деякими атрибутами і обробці саме цих атрибутів, та відповідних об'єктів [2]. Другий підхід, який власне і представляє цінність, заснований на протилежній ідеї: замість складних логічних правил використовується проста математична модель, – статистичний аналіз вже існуючих текстів. Початок цей підхід бере в роботах над методом LSA (Latent Semantic Analysis, неявний семантичний аналіз) [3]. Пізніше метод зазнав безліч модифікацій і отримав досить широку популярність. Сьогодні Google і ряд інших великих пошукових систем використовують один з параметрів даного методу (індекс tf * idf) при ранжуванні результатів [6].

Принцип роботи пошукового алгоритму згідно даного методу досить простий: чим частіше два слова зустрічаються в одних і тих же контекстах (документах), тим ближче вони за змістом.

Для LSA частота знаходження в конкретному документі розраховується якраз у вигляді індексу tf * idf, що розшифровується як «term frequency * inverse document frequency» [4]. Term frequency (частота терміна) – розраховується як кількість входжень конкретного терміна в конкретний документ, поділене на загальну кількість слів у цьому документі:

Document frequency (частота документа) – це кількість документів, в яких цей термін зустрічається, поділене на загальну кількість документів. Inverse document frequency, відповідно, це величина, зворотна document frequency, тобто idf = 1/df. Зазвичай, щоб пом'якшити ефект дії idf на загальний результат, замість самого значення береться його логарифм.

Відповідно, у загальному випадку частота появи терма, що є зворотною частотою документа (tf * idf модель), використовується для обчислення ваги d_i для терма і в документі (1):

$$d_i = tf_i * idf_i, \quad (1)$$

де tf_i є частотою появи терма і в документі, а idf_i є оберненою частотою появи терма і в усьому контексті.

Всі документи проходять ранжування відповідно до їх подібності введеному запиту. Відсутність спільних термінів у двох документах не обов'язково означає, що документи не схожі семантично. Аналогічно, релевантні введеному запиту документи можуть не містити такі терміни.

Як відомо, у рамках інформаційного пошуку зміст документів (наприклад веб-сторінок) є важливою характеристикою для аналізу та побудови оптимальної пошукової видачі результатів пошуку документів [2], оскільки вони не повинні містити екземпляри, контент яких дублюється на

інших сторінках; кількість інформаційного шуму має бути мінімальною, а основний контент – релевантним предмету пошуку. Отже, оцінка вебсторінок на предмет дублювання інформації та її новизни розглядається як необхідний етап при побудові оптимальних алгоритмів інформаційного пошуку. Для розв'язання даної задачі, що полягає у знаходженні дублікатів найкращим чином підходить метод винятковостей (синглів), – англ. single method [1], основна ідея якого полягає у розбитті текстів, що порівнюються, на вибрані з тексту послідовності слів (синглів), для кожного з яких обчислюється контрольна сума.

Міра близькості двох текстових документів $sim(D_i,D_j)$ визначалась на основі апарату умовних ймовірностей (2), а саме, як добуток ймовірності того, що випадкове слово *w* входить в документ D_i за умови, що воно входить в документ D_j, помножене на ймовірність входження цього слова в документ D_j.

$$sim(D_i,D_j) = P(w \in D_i \mid w \in D_j)P(w \in D_j) \quad (2)$$

В такому випадку, параметр новизни New_i документа D_i (3):

$$New_i = \frac{Rank_i * sim(D_i, PlusDisc)}{\log(i+1)\sum_{j=1}^{N} sim(D_i,D_j)}, \quad (3)$$

де N – загальна кількість веб-документів; D_j – j-й поточний документ; D_i – і-й документ; PlusDic – словник; $sim(D_i,D_j)$ – міра близькості документів і та j; $sim(D_i,PlusDic)$ – і-го документу та словника; $Rank_i$ – ранг і-го документу.

З точки зору підвищення достовірності оцінки релевантності документу до запиту доцільним буде використання вдосконаленого методу SeoRank [2] для визначення релевантності інформаційних блоків документу (веб-сторінки) щодо її основного змісту, який представлений на веб-сторінці у вигляді інформації у метатегах. Тобто відбуватиметься детальна оцінка складових документу. На відміну від існуючих раніше методів оцінки релевантності (наприклад, PageRank), вдосконалена форма SeoRank не розглядає релевантність інформаційних блоків відносно конкретних пошукових запитів і не враховує зовнішні параметри, такі як взаємозв'язки між ресурсами, фізичну доступність ресурсу, відповідність стандартам тощо, а дає можливість оцінити інформаційні блоки в межах конкретного документу (веб-сторінки) [2].

Формально модифікований SeoRank обчислюється як (4):

$$SeoRank = \sum_{i=1}^{4} a_i r_i, \quad (4)$$

де r_i – значення параметра; a_i – вага параметра; при чому сумарна вага (5):

$$\sum_{i=1}^{4} a_i = 1 \quad (5)$$

Відповідно, для обчислення модифікованого SeoRank використовуються наступні параметри [1]: 1) релевантність заголовку веб-сторінки («title») до тексту інформаційного блоку r_1 – відношення кількості входжень слів з заголовку у текст блоку до загальної кількості слів блоку; 2) релевантність ключових слів веб-сторінки («meta keywords») до тексту інформаційного блоку r_2 – відношення кількості входжень ключових слів у текст блока до загальної кількості слів блоку; 3) релевантність слів з опису веб-сторінки чи документа («meta description») до тексту інформаційного блоку r_3 – відношення кількості входжень слів з опису веб-сторінки у текст блоку до загальної кількості слів блоку; 4) релевантність заголовків веб-сторінки чи документа («headers») до тексту інформаційного блоку r_4 – відношення числа входжень слів з заголовків («H1»-«H6») веб-сторінки до загальної кількості слів з заголовків блоку.

Висновки і пропозиції. Виходячи з проведеного огляду сучасного стану досліджень в області оптимізації алгоритмів інформаційного пошуку встановлено наступні проблеми: велика кількість дубльованого контенту; відсутність розбиття результатів веб-пошуку за тематиками; значна кількість інформаційного спаму при перегляді документів, що значно впливає на час пошуку та перегляду документів. Відповідно, вимоги до швидкості пошуку, актуальності інформації з кожним днем зростають; водночас збільшуються і вимоги до методів та алгоритмів пошуку і подання інформації [1]. Неперервний розвиток інформаційного пошуку та зростання обсягів даних вимагає постійного покращення існуючих методів та розробку якісно нових підходів. Ці та інші фактори вказують на те, що проблема розробки та вдосконалення ефективних алгоритмів інформаційного пошуку у веб-системах є актуальною.

Відповідно заданим вимогам, перспективним в плані загальності та глибинності у рамках лінгвістичного аналізу текстів буде використання моделі векторного простору (VSM) [4]. Суть моделі полягає у математичному представленні текстів, в якій кожному документу зіставлений вектор, що виражає його зміст. Водночас, як необхідний етап при побудові оптимальних алгоритмів інформаційного пошуку розглядається оцінка вебсторінок на предмет дублювання інформації та її новизни. Для знаходження дублікатів найкращим чином підходить метод винятковостей (синглів), – англ. single method [1], основна ідея якого полягає у розбитті текстів, що порівнюються, на вибрані з тексту послідовності слів (синглів). З точки зору підвищення достовірності оцінки релевантності документу до запиту доцільним буде використання вдосконаленого методу SeoRank [2] для визначення релевантності інформаційних блоків веб-сторінки щодо основного змісту, який представлений на веб-сторінці у вигляді інформації у метатегах. Результати, отримані при проведенні даного дослідження можуть бути використанні при подальшому аналізуванні та вдосконаленні алгоритмів інформаційного пошуку.

Список літератури:

1. Ашманов И.С. Продвижение сайта в поисковых системах / И.С. Ашманов, А.А. Иванов – М.: Вильямс, 2016. – 304 с.
2. Колисниченко Д.Н. Поисковые системы и продвижение сайтов / Д.Н. Колисниченко – М.: Диалектика, 2014. – 272 с.
3. Крохина О.И. Первая книга SEO-копирайтера. Как написать текст для поисковых машин и пользователей / О.И. Крохина, М.Н. Полосина – М.: Инфра-Инженерия, 2012. – 216 с.
4. Маннинг К. Введение в информационный поиск / К. Маннинг, П. Рагхаван, Х. Шютце – М.: Вильямс, 2011. – 600 с.
5. Слабченко О.О., Сидоренко В.Н. Покращення якості первинних даних в задачах моделювання інтернет-співтовариств на основі комплексного застосування моделей сегментації, імпутації і збагачення даних // Вісник Кременчуцького національного університету імені Михайла Остроградського. – Кременчук: КрНУ, 2013. – Випуск 6(83). – С. 50-58.
6. Костенко П.П., Левченко І.В. Веб-сервіс уточнення релевантності веб-документів пошукової Видачі Google на основі поведінки користувача // Інженерні та освітні технології. Щоквартальний науково-практичний журнал – Кременчук: КрНУ, 2014. – Випуск 4(8). – С. 49-62.
7. Brin S., Page L. The anatomy of a large-scale hypertextual Web search engine // Computer Networks and ISDN Systems, 2004. – Pp. 107-117.

Терещенко В.В.
Кременчугский национальный университет
имени Михаила Остроградского
Терещенко В.Л.
Филия «Кременчугский институт
ВУЗ «Университет имени Альфреда Нобеля»

ПЕРСПЕКТИВЫ СОВЕРШЕНСТВОВАНИЯ ИНФОРМАЦИОННОГО ПОИСКА

Аннотация
В современных информационных технологиях сети Internet и поисковых машинах есть потребность в усовершенствовании поисковых алгоритмов. В рамках повышения эффективности поисковой выдачи предложено: использование модели векторного пространства (VSM); метод исключительностей (англ. single method). Для повышения достоверности оценки документа целесообразно использовать метод SeoRank. Результаты исследования будут полезны при усовершенствовании поисковых алгоритмов.
Ключевые слова: информационный поиск, поисковая система, поисковая выдача, релевантность, поисковый алгоритм, веб-страница.

Tereshchenko V.V.
Kremenchuk Mykhaylo Ostrohradskiy National University
Tereshchenko V.L.
Branch «Kremenchug Institute
of HEI «University of Alfred Nobel»

PROSPECTS FOR IMPROVING OF INFORMATION SEARCH

Summary
In modern information technologies and the Internet search engines there is a need to improve the search algorithms. As part of increasing efficiency within the search results proposed: the use of vector space model (VSM); method of exclusivity (Eng. single method). In terms of improving the reliability evaluation document emphasized the usefulness of the method SeoRank. The results will be useful in improving of search algorithms.
Keywords: information retrieval, search engine, search results, relevance; search algorithm, web page.