УДК: 004.738.52 DOI: 10.30929/1995-0519.2018.3.26-31

АНАЛІЗ СУЧАСНИХ МЕТОДІВ ІНФОРМАЦІЙНОГО ПОШУКУ

В. В. Терещенко
Кременчуцький національний університет імені Михайла Остроградського
вул. Першотравнева, 20, м. Кременчук, 39600, Україна. E-mail: darkwolfthehunter@gmail.com

У сучасних умовах розвитку інформаційних технологій мережі Internet та пошукових машин виникає потреба у нових методах забезпечення ефективного інформаційного пошуку. Проаналізовано сучасні методи інформаційного пошуку та, спираючись на вимоги сьогодення, виокремлено найважливіші аспекти. Відповідно, проаналізовано опубліковані у 2017 році фактори ранжування Google. У рамках підвищення ефективності пошукової видачі запропоновано: використання моделі векторного простору (VSM); використання методу винятковостей (англ. single method). Для підвищення достовірності оцінки документу наголошено на доцільності застосуваня методу SeoRank з метою визначення релевантності інформаційних блоків веб-сторінки щодо основного змісту, який представлений на веб-сторінці у вигляді інформації у метатегах. Результати, що отримані при проведенні даного дослідження можуть бути використанні при подальшому опрацюванні пошукових методик, розвитку засобів забезпечення пошукових систем, вдосконаленні пошукових алгоритмів.

Ключові слова: пошукова оптимізація, пошукова система, пошукова видача, інформаційний пошук.

АНАЛИЗ СОВРЕМЕННЫХ МЕТОДОВ ИНФОРМАЦИОННОГО ПОИСКА

В. В. Терещенко
Кременчугский национальный университет имени Михаила Остроградского
ул. Первомайская, 20, г. Кременчуг, 39600, Украина. E-mail: darkwolfthehunter@gmail.com

В современных условиях развития информационных технологий сети Internet и поисковых машин возникает потребность в новых методах обеспечения эффективного информационного поиска. Проанализированы современные методы информационного поиска и, опираясь на требования современности, выделены важнейшие аспекты. Соответственно, были проанализированы опубликованные в 2017 году факторы ранжирования Google. В рамках повышения эффективности поисковой выдачи предложено: использование модели векторного пространства (VSM); использование метода исключительности (англ. single method). Для повышения достоверности оценки документа отмечено целесообразность применения метода SeoRank с целью определения релевантности информационных блоков веб-страницы относительно основного содержания, который представлен на веб-странице в виде информации в метатегах. Результаты, полученные при проведении данного исследования могут быть использованы при дальнейшей обработке поисковых методик, развития средств обеспечения поисковых систем, совершенствовании поисковых алгоритмов.

Ключевые слова: поисковая оптимизация, поисковая система, поисковая выдача, информационый поиск.

АКТУАЛЬНІСТЬ РОБОТИ. 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 сучасної економіки, а й науки та технологій.

Пошукова машина розвивається в різних напрямках: з'являються нові чинники ранжування або змінюється їх пріоритет, посилюються вимоги до якості сайтів та їх посилальних зв'язків (з'являються нові антиспам-алгоритми), змінюється сам формат взаємодії пошуку з користувачем, а також з'являються нові сервіси, що спрощують пошук інформації.

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, відповідно, проблема вдосконалення методів інформаційного пошуку є актуальною.

Мета роботи - є суттєве поліпшення результатів інформаційного пошуку за показниками релевантності.

МАТЕРІАЛ І РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ. 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) [2].

В рамках дослідження проблем інформаційного пошуку працювали багато науковців: 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 методів інформаційного пошуку не розв'язана повністю та залишається актуальною.

Відповідно до загальних принципів організації інформаційного пошуку, в основі кожного пошукового методу, його алгоритмів 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.

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 та методів.

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 [11] 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 A визначалася вагою посилання, переданою сторінкою B. Таким чином, PageRank був методом обчислення ваги сторінки шляхом підрахунку важливості посилань на неї [3].

У світлі останніх заходів з боку відомих пошукових систем, прийнятих для боротьби з покупними посиланнями, накруткою та іншими маніпулятивними способами, які б призводили до штучного підвищення рейтингу того чи іншого веб-ресурсу в пошуковій видачі, значно підвищилася роль так званих «поведінкових факторів» як елементів просування сайту в ТОП пошукової видачі.

Поведінкові фактори – це показники, які характеризують роботу користувача з пошуковою видачею та безпосередню його поведінку на сайті. Їхнє головне завдання полягає у покращенні якості побудови пошукової видачі [1].

Вплив поведінкових факторів на ранжування в пошукових системах незаперечний. До основних поведінкових факторів належать [5]:

1. Кліки у видачі пошукової системи. Основне значення надається першому і останньому клікам, адже, згідно алгоритмів Яндексу, зазвичай саме вони вважаються найбільш доречними.

2. Відвідуваність ресурсу, яка вказує на його популярність і затребуваність на просторах мережі.

3. Час знаходження як на сайті в цілому, так і в його окремих розділах зокрема. Це найважливіший критерій оцінки якості ресурсу, адже хороший портал завжди затримає відвідувачів на тривалий термін в той час як поганий користувачі залишають практично відразу.

4. Глибина перегляду, яку обчислюють з кількості переглянутих сторінок. Цей фактор залежить від попереднього. Чим більше часу користувач провів на сайті, тим більше сторінок було переглянуто, тим якіснішим будуть вважати ресурс пошукові системи.

5. Показник повернень – кількість користувачів, які формують постійну цільову аудиторію.

6. Показник відмов, що охоплює користувачів, які не стали переглядати на сайті більше однієї сторінки. Цей фактор вважається негативним, вказуючи на низьку якість сайту і нерелевантність запиту. Важливо звернути увагу також на те, що цей критерій не може вважатися основним, адже відвідувачі могли покинути сторінку не тільки тому, що вона не відповідала їхнім очікуванням, а і внаслідок того що вони відразу знайшли відповідь на своє питання.

7. Посилання на сайт. Кількість переходів на веб ресурс не лише з пошукових видач а й з інших джерел (наприклад соціальні мережі і т.д.).

Зазвичай власниками сайтів проводиться аналіз поведінкових факторів за допомогою систем веб-аналітики що під'єднуються безпосередньо до сайту. До найпопулярніших серед них відносяться Яндекс.Метрика та Google Analytics [4]. Цей факт необхідно враховувати, обираючи принципи за якими буде працювати пошуковий алгоритм. Однак, при використанні власником веб-ресурсу декількох лічильників одночасно (наприклад Google Analytics та Яндекс.Метрика), помітна різниця у підрахунках статистики.

Вагомі причини таких відмінностей полягають у тому, що системи аналітики оперують різними даними й рахують одні й ті ж показники по-різному. Яндекс.Метрика та Google Analytics можуть відбражати різні дані з різних причин [1]:

1. Лічильники встановлеі в різних місцях HTML-коду. Наприклад, якщо лічильник аналітики встановлено в тегові <head>, а метрики перед закриваючим тегом </body>, то більшість користувачів, що не дочекались завантаження сторінки, не відобразяться в статистиці.

2. Неправильні налаштування часової зони. В Яндекс.Метрика та Google Analytics існує можливість в налаштуваннях лічильника задати часовий пояс для розрахунку статистики. Якщо вказано різні часові пояси, то статистичні дані будуть відрізнятися.

3. Налаштування фільтрів. Різні вказівки при налаштуванні фільтрів привзодять до різниці у відображуваних даних.

Крім того, відмінності полягають й у розумінні термінології [4]. Звичайний користувач звик вважати візити рівноцінними відвідуванню, однак це не є правильним. Візити в Яндексі – це кількість сеансів взаємодії користувачів з сайтом, під час якого переглядаються одна чи більше сторінок. Візит завершується через 30 хвилин після відсутності активності з боку користувача. До сеансу відвідування в Google відносять всі дані щодо використання (перегляди, транзакції й т.п.).

Та ж сама ситуація й з показником відмов. В Google це процент відвідувань, в ході яких було відкрито не більше одної сторінки. В Яндексі це доля визитів, в рамках яких відбувся лише один перегляд сторінки.

Однак, при визначенні релевантності пошукові системи в першу чергу звертають увагу на те, скільки разів на сторінці зустрічається фраза, тотожна запиту користувача. Цей параметр називається частотою ключового слова. Чим він вищий, тим релевантнішим вважається сайт. Донедавна оптимізатори спеціально збільшували частоту ключових слів аж до повної нечитаності текстів. На даний момент пошукові системи активно борються з подібними методами і знижують ранжування при їх виявленні.

Для визначення частоти ключових слів використовуються спеціальні математичні алгоритми, що вираховують кількість входжень запитів на обсяг обумовленого тексту. При цьому оптимальним співвідношенням вважається 3-5%. Так як роботи пошукових систем не здатні оцінювати тексти з погляду читабельності, ця обставина і дозволяє оптимізаторам збільшувати частоту ключових слів до певної межі, що, з одного боку, порушує правила використання пошукових систем, а з іншого – не виходить за рамки встановлених ними критеріїв .

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 di 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 веб-сторінок 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; Dj – j-й поточний документ; D_i – i-й документ; PlusDic – словник; $sim(D_i,D_j)$ – міра близькості документів і та j; $sim(D_i,PlusDic)$ – i-го документу та словника; $Rank_i$ – ранг i-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 r1 – 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 r2 – 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 r3 – 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 r4 – відношення числа входжень слів з заголовків («H1»-«H6») веб-сторінки до загальної кількості слів з заголовків блоку.

ВИСНОВКИ. В результаті аналізу опублікованих у 2017 році факторів ранжування Google можна дійти до наступних висновків відносно принципів роботи сучасних пошукових машин:

1. Збільшується увага до релевантного, якісного контенту («Relevant, holistic content is more important than ever») та вплив поведінкових факторів.

2. Зменшується вплив ключових слів. («Keywords are becoming increasingly obsolete»).

3. Mobile Friendly: створення робочої мобільної версії сайту – обов'язковий крок у просуванні ресурсу. Вже з 2015 року Google почав ранжувати вище сайти, адаптовані під мобільні пристрої.

4. Роль соціальних мереж зростає – якщо є переходи з соцмереж, значить, сайт матиме більший рейтинг в пошуковій видачі («Social signals – a bonus for positive rankings»).

Виходячи з проведеного огляду сучасного стану досліджень в області оптимізації методів та 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 методів 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 веб-сторінок 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 методів та алгоритмів інформаційного пошуку.

ЛІТЕРАТУРА

1. Ашманов И. С., Иванов А. А. Продвижение сайта в поисковых системах. М.: Вильямс, 2016. 304 с.

2. Колисниченко Д. Н. Поисковые системы и продвижение сайтов. М.: Диалектика, 2014. 272 с.

3. Крохина О. И., Полосина М. Н. Первая книга SEO-копирайтера. Как написать текст для поисковых машин и пользователей. М.: Инфра-Инженерия, 2015. 236 с.

4. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. М.: Вильямс, 2017. 640 с.

5. Fetterly D., Manasse M., Najork M. Spam, Damn Spam, and Statistics. *Int'l Workshop on the Web and Databases, ACM Press,* 2004. PP. 1–6.

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. Вип. 4 (8). С. 49–62.

7. Славко О. Г., Костенко П. П. Забезпечення якості обслуговування інформаційних систем на основі локальної моделі керованого процессу. *Радіоелектронні і комп'ютерні системи.* 2012. Вип. 1 (53). С. 99–104.

8. 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. Вип. 6 (83). С. 50–58.

9. Заїка А. В., Філенко М. І. , Остапченко А. С., Григорова Т.А. Моделювання архітектурних рішень підтримки мультисайтовості для організації інформаційних систем. *Вісник Кременчуцького національного університету імені Михайла Остроградського.* Кременчук: КрНУ, 2015. Вип. 3 (92), ч.1. С. 54–59.

10. Терещенко В. В., Терещенко В. Л. Перспективність вдосконалення систем інформаційного пошуку. *Четверта Всеукраїнська науково-практична конференція «ІТ-Перспектива»*. Кременчук: КрНУ, 2017. С. 26–28

11. Alexandros N., Mark M. Detecting Spam Web Pages through Content Analysis. *Microsoft Research,* 2012. PP. 1–6.

12. Brin S., Page L. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems,* 2004. PP. 107–117.

13. Ganz A., Sieh L., Behavioral factors and SEO. *Proceedings of 24th International Conference on Computer Communications and Networks (ICCCN 2015),* Las Vegas, Nevada, USA August 3 - August 6, 2015, Scottsdale, Arizona, USA. PP. 218–223.

THE ANALYSIS OF CURRENT METHODS OF INFORMATION SEARCH

V. Tereschenko
Kremenchuk Mykhaylo Ostrohradskiy National university
vul. Pershotravneva, 20, Kremenchuk, 39600, Ukraine. E-mail: darkwolfthehunter@gmail.com

Purpose. To analyze current methods of information search. In modern conditions of development of information technologies Internet and the search engine there is a need for new methods to ensure effective information search. The search engine develops in different directions: new ranking factors appear or their priority changes, requirements to the quality of sites and their referral links increase (new anti-spam algorithms appear), changed the format of the interaction of the search with the user, there are new services that make searching easier. **Methodology.** In this paper modern methods of information search were investigated, the published ranking factors of Google published in 2017 were analyzed, and the prospect of using the vector space model (VSM) was considered. **Results.** The analysis of published in 2017 ranking factors Google can come to the following conclusions: 1. It is necessary to increase attention to relevant, quality content («Relevant, holistic content is more important than ever») and the influence of behavioral factors; 2. Reduces the impact of keywords. («Keywords are becoming increasingly obsolete»); 3. Mobile Friendly: establishment of a working mobile version of the site – a mandatory step in promoting resource. Starting with 21 th April 2015 Google ranked above sites adapted to mobile devices; 4. The role of social networks is growing - if there are referrals from social networks, then the site will have a greater ranking in search engine results page («Social signals – a bonus for positive rankings»). Accordingly, the use of vector space model (VSM) will be perspective in terms of generality and depth in terms of linguistic analysis of texts.Web pages must be evaluated for the finding purpose of duplication of information and its novelty. To find duplicates, the best method is the method of exclusiveness (singles) – «single method». From the point of view of increasing the reliability of the assessment of relevance of the document to the request, it will be expedient to use the improved SeoRank method. **Originality.** For the first time, it was analyzed the factors of Google ranking from the standpoint of the search engine and not the SEO Optimizer; based on the requirements of the present, the most important aspects are outlined. Within the framework of the analysis, the most adapted methods are identified that are acceptable for improving the search engine. **Practical value.** The results obtained during this study can be used for further analysis and improvement of methods and algorithms of information search. References 13.

Key words: search engine optimization, search engine, search engine results, information search.

REFERENCES

1. Ashmanov, I. S., Ivanov A. A. (2016), *Prodvigenye sayta v poiskovyh sistemah* [Website promotion in search engines], Willyams, Moscow, Russia.

2. Kolosnichenko, D. N. (2014), *Poiskovye sistemy i prodvigenye saytov* [Search engines and website promotion], Dialektika, Moscow, Russia.

3. Krohina, O. I., Polosina M. N. (2015), *Pervaya kniga SEO-kopyritera* [First book of SEO copyriter], Infra-Ingeneria, Moscow, Russia.

4. Manning, K., Raghavan, P., Shutce, H. (2017), *Vvedenie v infirmacionnyi poisk* [Introduction into information search], Willyams, Moscow, Russia.

5. Fetterly D., Manasse M., Najork M. Spam, Damn Spam, Statistics (2004), *Int'l Workshop on the Web and Databases, ACM Press*, pp. 1-6.

6. Kostenko, P. P., Levchenko, I. V. (2014), *Web-servis utochnennya relevantnosti web-dokumentiv poshukovoi vidachi Google na osnovi povedinki koristuvacha* [Web-service for clarification relevant web-documents of search results of Google based on user behavior], Inzhenerni ta osvitni tehnologii, Vip. 4 (8), PP. 49-62, Ukraine.

7. Slavko, O. G., Kostenko, P. P. (2012), *Zabezpechennya yakosti obsluhovuvannya informatsiynikh system na osnovi lokal'noyi modeli kerovanoho protsesa* [Quality of Service providing of information systems on a base of a controlled process local model], Radioelektronni y komp'yuterni systemy, Vip. 1 (53), PP. 99-104, Ukraine.

8. Slabchenko, O. O., Sidorenko V. N. (2013), *Pokrashhennya yakosti pervinnih danih v zadachah modelyuvannya internet-spivtovaristv na osnovi kompleksnogo zastosuvannya modelej segmentacii, imputacii i zbagachennya danih* [The improvement of initial data quality in modeling ploblems of online communities on the base of combined implementation of segmentation, imputation and data enrichment models], Visnik Kremenchuckogo nacionalnogo universitetu imeni Mykhaila Ostrogradskogo, Vipusk 6 (83), PP. 50-58, Ukraine.

9. Zaika, A. V., Filenko, M. I., Ostapchenko, A. S., Hryhorova, T. A. (2015), *Modelyuvannya arhitekturnih rishen pidtrimki multisajtovosti dlya organizacii informaciynih sistem* [Design of architecture for support multi-site in information systems], Visnik Kremenchuckogo nacionalnogo universitetu imeni Mykhaila Ostrogradskogo, Vip. 3 (92) part 1, PP. 54-59.

10. Tereschenko, V. V., Tereshchenko, V. L. (2017), *Perspektivnist vdoskonalennya sistem informatsiynogo poshuku* [The prospectivity of improving information search systems], IV Vseukrainska naukovo-praktychna konferentsiya «IT-Perspektiva», PP. 26-28.

11. Alexandros, N., Mark, M. (2012), Detecting Spam Web Pages through Content Analysis. *Microsoft Research*, pp. 1-6.

12. Brin, S., Page, L. (2004), The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, pp. 107-117.

13. Ganz, A., Sieh, L., (2015), Behavioral factors and SEO, *Proceedings of 24th International Conference on Computer Communications and Networks (ICCCN 2015)*, Las Vegas, Nevada, USA August 3 – August 6, 2015, Scottsdale, Arizona, USA. pp. 218-223.

Стаття надійшла 03.05.2018.