УДК 025.4.03

С.В. Знахур, О.Ю. Мізяк

Харківський національний економічний університет, Харків

АЛГОРИТМИ ПОШУКУ РЕЛЕВАНТНИХ ДОКУМЕНТІВ У ІНФОРМАЦІЙНИХ МЕРЕЖАХ

Äîñë³äæåíî ìîæëèâîñò³ ïîáóäóâàííÿ ³íôîðìàö³éíî-ïîøóêîâî¿ ñèñòåìè, à ñàìå ¿¿ àëãîðèòì³â ³íäåêñàö³¿, ïîøóêó ³ ðàíæèðóâàííÿ äàíèõ. Ðîçãëÿíóò³ àëãîðèòìè êîíêóðåíòíèõ ïîøóêîâèõ ñèñòåì, ¿õ ïåðåâàãè òà íåäîë³êè. Áóâ ïðåäñòàâëåíèé ìåòîä, ÿêèé ïîëÿãàº â ñòâîðåíí³ ìåõàí³çìó, ïîáóäîâàíîìó íà òðüîõ àëãîðèòìàõ ïîøóêó. Àëãîðèòì çâàæåííîãî çíîííîãî ðàíæèðóâàííÿ ï³äâèùóº êîåô³ö³ºíò ðåëåâàíòíîñò³ äëÿ äîêóìåíò³â, â ÿêèõ òåðì³íè çàïèòó çíàõîäÿòüñÿ ó âèçíà÷åíèõ çîíàõ. Àëãîðèòì ðàíæèðóâàííÿ íà îñíîâ³ ÷àñòîò âõîäæåííÿ òåðì³í³â ðîçðàõîâóº êîåô³ö³ºíò ðåëåâàíòíîñò³ â çàëåæíîñò³ â³ä ÷àñòîòè ïîÿâè òåðì³íó çàïèòó ó äîêóìåíò³ òà ó âñ³é êîëåêö³¿. Àëãîðèòì áëèçüêîñò³ òåðì³í³â ïîëÿãàº ó â³äøóêàíí³ íàéá³ëüø ðåëåâàíòíèõ òåðì³í³â îäèí äî îäíîãî. Ïîºäíàííÿ äàíèõ àëãîðèòì³â çàáåçïå÷óº ï³äâèùåííÿ åôåêòèâíîñò³ ïîøóêó íà ð³çíèõ ³íôîðìàö³éíèõ çàïèòàõ.

Ключові слова: *ранжирування , релевантність, індексація, термін, пошукова система, документ, лексема, алгоритм.*

Вступ

Інформаційний пошук організовується за допомогою інформаційно-пошукових систем – комплексів окремих апаратів пов'язаних між собою і призначених для отримання інформації в колекції документів, яка відповідає заданому інформаційному запиту [1 – 3]. Головним завданням пошукової системи є необхідність надавати релевантні результати на запити користувача. Доцільно відзначити, що результати повинні відповідати принципу «максимальна якість мінімальна кількість [4]. Для адаптації системи до таких запитів, які відповідають принципам її роботи, створюються алгоритми пошукових систем. Для того, щоб задовольнити зростаючим потребам користувачів, пошукові системи налічують та постійно вдосконалюють алгоритми пошуку, додають нові функції і можливості, та прискорюють його роботу [5].

Огляд існуючих методів та алгоритмів. У світі існує пріорітет використання різних пошукових систем по відношенню до регіонів світу та їх ринків інформаційних запитів.

Для переважної більшості «вітчизняних» пошукових систем процес індексації та пошуку інформації є регіональним. Тобто, основна відмінність від «зарубіжних» систем в тому, що ресурси які індексуються, розташовані в доменних зонах, де домінує російська мова. Це створює лінгвістичний бар’єр між зарубіжними інформаційними джерелами, та вітчизняним ринком пошукових запитів [6, 7].

На рис. 1 відображена статистика Інтернет-сервісу Bigmir.net, яка характеризує розподіл ринку пошукових запитів в Україні за 2012-й рік.

Відомий механізм інформаційно-пошукової системи «Yandex» будується на алгоритмі побудови релевантних результатів з оглядом на морфологію російської мови. Морфологічний розбір не прив'язаний до словника – в випадку відсутності терміна в словнику, знаходяться найбільш схожі, і по ним будується модель словозміни.

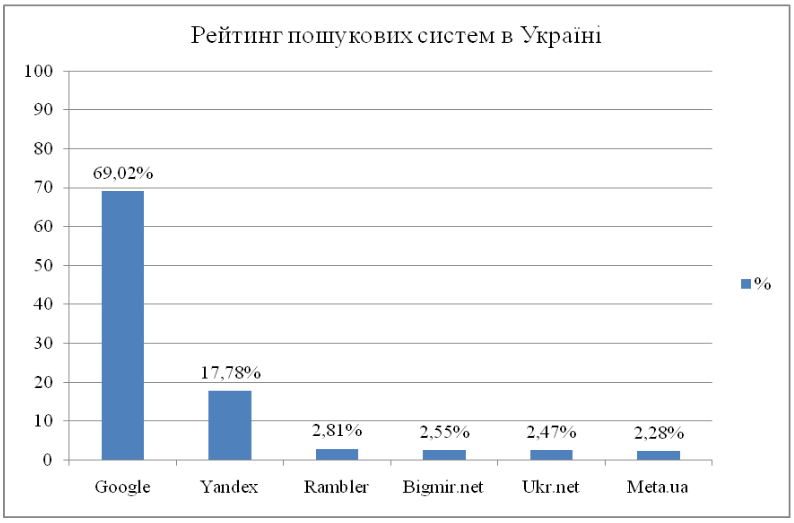

Рис. 1. Рейтинг найпопулярніших глобальних пошукових систем в Україні

Недоліком цього методу є низька пошукова здатність на мовах відмінних від російської та залежність результату запиту від регістра його набору. Це не дозволяє отримувати релевантні відповіді на пошукові запити і в свою чергу, не дає можливості виходу системи на зарубіжний ринок. Даний недолік свідчить про відсутність універсальності пошукового алгоритму.

Найбільш близьким по сукупності ознак до запропонованого нами, є метод «Page Rank» пошукової системи «Google» в якому релевантність документа визначається на основі інформації про документи, які посилаються на нього, залежить від їх числа, а також релевантності цих посилань. Такий підхід дозволяє забезпечити видачу якісних результатів.

Недоліком цього методу є лінійне збільшення інформаційного «шуму» в залежності від збільшення кількості посилань. Це обумовлено тим, що документи можуть посилатися на ресурси з застарілою або видаленою інформацією, які не оновлюються власником. Це створює ризик отримати відповіді, які не будуть повністю задовольняти умовам запиту.

Актуальність дослідження. Актуальність поставленого дослідження обумовлена існуванням певних проблем у напрямку інформаційного пошуку, а саме:

- зріст інформаційного контенту, який супроводжується зростом інформаційного шуму та багаторазовим дублюванням інформації [8];
- низька структурованість даних;
- відсутність єдиної системи стандартизації та концепції розвитку Інтернет та відсутність центра управління розвитком глобальної мережі;
- постійна зміна технологій зберігання та представлення даних;
- зростання вимог з боку користувачів пошукових систем.

Формулювання мети статті. Метою даної статті є дослідження можливостей пошуку і ранжирування даних для надання найбільш якісного, релевантного результату. Необхідно розробити такі алгоритми пошуку, які могли б наслідувати можливості алгоритмів конкурентних систем, а також внести нові ідеї, тим самим зробивши пошукову систему унікальною. Основне вирішення даної проблеми полягає у впровадженні методів інтелектуальної обробки інформації.

Мета роботи пошукового механізму полягає в організації пошуку релевантної інформації в інформаційних мережах за довільним запитом.

Поряд з цим, метою системи є пошук таких документів колекції , які є найбільш релевантними по відношенню до довільних інформаційних потреб користувача. Документ називається релевантним, якщо, з точки зору користувача, він містить коштовну інформацію, яка задовольняє його інформаційні потреби.

Основна частина

Функціонування механізму пошукової системи можна поділити на два основні, незалежні один від одного завдання: індексація метаданих отриманих від пошукового агента і організація пошуку на підставі запиту користувача і індексованих в системі документів. Для отримання виграшу у швидкості пошуку, проводиться індексація ресурсів і на їх підставі виконується ранжирування інформації. Процес індексації складається з наступних етапів:

- збір документів, які підлягають індексації;
- представлення тексту у вигляді лексем;
- попередня обробка інформації (стеммінг, лематизація);
- індексація документа.

Індексація документу. У колекції кожен документ має власний ідентифікатор. У процесі побудови індексу даний ідентифікатор або присвоюється новим документом, або оновлюється, якщо документ вже існує. Вхідною інформацією для індексування є набір нормалізованих лексем для кожного документа, який розглядається як список пар «термін-документ».

Основним етапом в процесі індексування є сортування списку термінів, в результаті чого вони розташовуються в алфавітному порядку. Після цього зразки одного і того ж терміну групуються, а результат поділяється на таблицю атрибутів і таблицю словосполучень. Таким чином індексації документ стає придатним для проведення пошуку.

Алгоритми пошуку. Нами будо розроблено та розглянуто три алгоритми пошуку які у поєднанні між собою підвищують його ефективність. Перший з них алгоритм зваженого зонного ранжирування .

Веб-документи супроводжуються метаданими, які кодуються у вигляді, доступному для розпізнавання сервером. Під метаданими представляється неоднорідна інформація про терміни документа, яка містить сукупність ознак таких як: присутність терміна в заголовку сторінки або мета-тегах, чи є термін посиланням на інший документ, жирним, курсивним, підкресленим тощо.

Такі ознаки називаються зонами і дозволяють збільшити або зменшити важливість терміна в порівнянні з іншими термінами оброблюваного документа. Така підтримка обробки запитів полягає в створенні таблиці атрибутів для індексованих документів яка вказує на приналежність терміна до тієї чи іншої зони. Даний підхід дозволяє забезпечити алгоритм зваженого зонного ранжирування .

Даний алгоритм привласнює для пари документа (d) та запита (q) значення релевантності на відрізку [0..1], обчислюючи лінійну комбінацію зонних показників, до якої кожна зона документу вносить булеве значення. Нами було розглянуто колекцію документів, кожен з яких має l-зон. так що:

$$\sum_{i=1}^{l} g_i = 1 \qquad (1)$$

де g_i – коефіцієнти зон а також $g_i \in [0.1]$.

Зважаючи на цю умову зважену зонну релевантність розраховують по формулі:

$$\sum_{i=1}^{l} g_i * s \qquad (2)$$

де s – булева величина, що означає відповідність (або її відсутність) між запитом (q) і і-ою зоною.

Ваги g_1 .. g_i вказуються експертами або користувачем. Однак набагато частіше ваги визначаються на основі навчальних прикладів, оцінених заздалегідь. Цей метод відноситься до загального класу методів ранжирування в інформаційному пошуку під назвою «методи ранжирування на основі машинного навчання».

Наступним алгоритмом є ранжирування на основі частот входження термінів.

При використанні зваженого зонного ранжирування ранг документа залежить від наявності термінів запиту в зонах документа. Наступним кроком є визначення частот входження термінів у документах: документ, де термін запиту зустрічається частіше, слід вважати більш релевантним і привласнити йому більш високе значення релевантності. Це обґрунтовується тим, що терміни оброблюються інтерфейсом пошукової машини у вільному вигляді, без сполучних операторів.

Такий стиль, розглядає запит як безліч слів. Отже, для підрахунку показника документа достатньо підсумовувати показники його відповідності кожному з слів запиту. Для цього кожному терміну, виявленому в документі привласнюється вага, що залежить від кількості поява цього терміна в даному документі. Для оцінки відповідності між терміном запиту і документом використовується три схеми зважування.

Перша з них полягає в тому, що вага терміна дорівнює кількості входжень його до документу. Ця схема зважування називається частотою терміна і позначається як tf_{td}, де індекс (t) позначає термін, а індекс (d) - документ. В рамках цієї схеми точний порядок термінів у документі ігнорується, а основна увага приділяється кількості входжень кожного терміна до документу. Для коригування значимості термінів документа між собою застосовується модифікація обчислення частоти терміна названа сублінійним маштабуванням. Вона полягає у використанні логарифма до частоти термінів:

$$wf = \begin{cases} 1 + \log(tf), tf > 0 \\ 0, tf \le 0 \end{cases}, \qquad (3)$$

де tf – частота терміну у документі, wf – сублінійна частота.

Представлена модифікація дозволяє зменшити важливість найбільш повторюваних термінів.

Така схема зважування має серйозний недолік: при ранжируванні документа за запитом всі терміни вважаються однаково важливими. Насправді, деякі терміни мають малу або нульову розрізняльну силу при визначенні релевантності. Для того щоб усунути зазначений недолік, вводиться механізм послаблення впливу «популярних» термінів. Для цього використовується друга схема зважування яка полягає у використанні документної частоти - кількості документів в колекції, що містять термін.

Це обумовлено тим, що, намагаючись знайти відмінності між документами з метою їх ранжирування за запитом, доцільніше використовувати статистичні показники самих документів (наприклад, кількість документів, що містять заданий термін), ніж статистичні показники колекції в цілому. Для коригування ваги терміна з використанням документної частоти необхідно визначити зворотну документну частоту за формулою:

$$idf = \log \frac{N}{df}, \qquad (4)$$

де N – загальна кількість документів в колекції, df – документна частота, idf – зворотня документна частота.

Найбільш ефективною схемою зважування є комбінація частоти терміна в документі і зворотної документної частоти. Дана схема привласнює кожному терміну в документі його коефіцієнт на підставі формули:

$$wf = idf = wf_{t,d} * idf_t, \qquad (5)$$

де wf – сублінійна частота для терміна (t) у документі (d), idf – зворотна документна частота, wf-idf – коефіцієнт терміна у документі.

Вага wf-idf має такі властивості:

- досягає максимального значення, якщо термін зустрічається багато раз в невеликій кількості документів (тим самим посилюючи їх відмінність від інших документів);
- зменшується, якщо термін зустрічається в документі лише кілька разів або зустрічається в багатьох документах (тим самим формуючи менш виражений сигнал про релевантність документа);
- досягає мінімального значення, якщо термін зустрічається практично у всіх документах [9].

Після визначення ваг всіх термінів запиту в усіх існуючих документах визначається міра перекриття: релевантність всіх документів дорівнює сумі входжень всіх термінів запиту в цей документ:

$$S(q,d) = \sum_{t \in d} (wf - idf), \qquad (6)$$

де wf-idf – коефіцієнт терміну у документі, S(q,d) – сума коефіцієнтів термінів запиту (q) на документ (d).

Релевантним документом буде вважатися, той документ, в якому сума коефіцієнтів релевантності для всіх термінів запиту буде найбільшою.

У випадку, коли запит складається більше ніж з одного терміну, доцільно організовувати пошук не тільки по частотним, та атрибутним характеристикам кожного терміну, а також з урахуванням фраз та словосполучень. Це дозволяє підвищити якість пошуку. Ідея алгоритму полягає у знаходженні рівня близькості термінів один до одного і розрахунку радіусу близькості. Пари термінів документу, які мають радіус – одиницю, представляють максимальний рівень близькості. Для кожної такої пари розраховується частота словосполучення і методика розрахунку коефіцієнту релевантності зводиться до використання частотних характеристик усіх словосполучень введеного запиту.

Поєднання коефіцієнтів усіх словосполучень виконується за формулою:

$$S(qc, d) = \sum_{i=0}^{k} (tfc(t_i + t_{i+1}), \qquad (7)$$

де k – кількість термінів у запиті, t – термін, tf – частота комбінації термінів, qc – документ, який містить комбінації слів, d – документ.

Релевантним документом буде вважатися, той документ, в якому сума коефіцієнтів релевантності для всіх можливих словосполучень запиту буде найбільшою.

Процес збільшення радіусу близькості термінів, з одного боку повинен покращити якість пошуку для великих запитів у ситуаціях, коли необхідне саме пошук словосполучень, а з іншого боку погіршує її при коротких запитах.

Поєднання цих алгоритмів забезпечується шляхом об'єднання коефіцієнтів по кожному алгоритму та визначення значимості кожного алгоритму для різних інформаційних запитів.

Висновки

Запропоновані нами алгоритми сприяють універсальності організації пошуку. Представлений підхід дозволяє працювати з різними наборами вхідних даних, за допомогою приведення цих даних до єдиного стандарту. Використання пошуку по словосполученням вирішує проблему фразових запитів, завдяки знаходженню пар термінів у документі. Для кожної такої пари розраховується частота появи словосполучення у документі і на основі цих даних знаходиться перелік релевантних документів.

Список літератури

1. *Ландэ, Д.В. Навигация в сложных сетях: модели и алгоритмы / Д.В. Ландэ, А.А. Снарский, И.В.. Безсуднов. – М.: Либроком, 2009. –264 с.*

2. *Мізяк,, О.Ю. Розроблення пошукового механізму для глобальних мереж / О.Ю. Мізяк // "Збірник наукових праць студентів спеціальностей "Інформаційні управляючі системи і технології", "Комп'ютерний еколого-економічний моніторинг"" [Текст]: /редкол.: В.С. Пономаренко [та ін.]. – Харків: ХНЕУ, 2011. – 346 с.*

5. *Ландэ, Д. В. Поиск знаний в Internet / Д. В. Ландэ. - М. : «Диалектика», 2005. –272 с.*

6. *Ашманов, И. С. Продвижение сайта в поисковых системах / И. С. Ашманов. - М. : «Вильямс», 2007. - 304 с.*

7. *Кадеев, Д.Н. Информационные технологии и ýëåê-òðîííûå êîììóíèêàöèè / Ä.Í. Êàäååâ.- Ì.: Ýëåêòðî, 2005. – 250 с.*

8. *Колисниченко, Д.Н. Поисковые системы и продвижение сайтов в Интернете / Д.Н. Колисниченко. - М.: Диалектика, 2007. – 272 с.*

9. *Маннинг К. Введение в информационный поиск К. Маннинг, П. Рагхаван, Ч. Шютце – М., 2011. – 528 с.*

Надійшла до редколегії 3.10.2012

Рецензент: д-р. техн. наук, С.В. Лістровой, Національний аерокосмічний університет ім. Жуковського «ХАІ», Харків.

АЛГОРИТМЫ ПОИСКА РЕЛЕВАНТНЫХ ДОКУМЕНТОВ В ИНФОРМАЦИОННЫХ СЕТЯХ

С.В. Знахур, А.Ю. Мизяк

Èññëåäîâàíû âîçìîæíîñòè ïîñòðîåíèÿ èíôîðìàöèîííî-ïîèñêîâîé ñèñòåìû, à èìåííî åå àëãîðèòìîâ èíäåêñàöèè, ïîèñêà è ðàíæèðîâàíèÿ äàííûõ. Ðàññìîòðåíû àëãîðèòìû êîíêóðåíòíûõ ïîèñêîâûõ ñèñòåì, èõ ïðåèìóùåñòâà è íåäîñòàòêè. Áûë ïðåäñòàâëåí ìåòîä, êîòîðûé çàêëþ÷àåòñÿ â ñîçäàíèè ìåõàíèçìà, ïîñòðîåííîì íà òðåõ àëãîðèòìàõ ïîèñêà. Àëãîðèòì âçâåøåííîãî çîííîãî ðàíæèðîâàíèÿ ïîâûøàåò êîýôôèöèåíò ðåëåâàíòíîñòè äëÿ äîêóìåíòîâ, â êîòîðûõ òåðìèíû çàïðîñà íàõîäÿòñÿ â îïðåäåëåííûõ çîíàõ. Àëãîðèòì ðàíæèðîâàíèÿ íà îñíîâå ÷àñòîò âõîæäåíèÿ ñëîâà ðàññ÷èòûâàåò êîýôôèöèåíò ðåëåâàíòíîñòè â çàâèñèìîñòè îò ÷àñòîòû ïîÿâëåíèÿ òåðìèíà â äîêóìåíòå è âî âñåé êîëëåêöèè. Àëãîðèòì áëèçîñòè òåðìèíîâ çàêëþ÷àåòñÿ â íàõîæäåíèè íàèáîëåå ðåëåâàíòíûõ òåðìèíîâ äðóã ê äðóãó. Ñî÷åòàíèå äàííûõ àëãîðèòìîâ îáåñïå÷èâàåò ïîâûøåíèå ýôôåêòèâíîñòè ïîèñêà íà ðàçëè÷íûõ èíôîðìàöèîííûõ çàïðîñàõ.

Ключевые слова: *ранжирование, релевантность, индексация, термин, поисковая система, документ, лексема, алгоритм.*

THE SEARCH ALGORITHM RELEVANT DOCUMENTS IN INFORMATION SYSTEMS

S.V. Znakhur, O.Y. Miziak

Investigated the possibility of building an information retrieval system, namely, its algorithms indexing, searching and ranking data. The algorithms of the search of competitive systems, their advantages and disadvantages. Presented method, which is to create a mechanism, based on three search algorithms. Weighted zone ranking algorithm improves overall relevance to the documents in which the query terms are in certain areas. Ranking algorithm is based on the frequency of the occurrence date to calculate the relevance depending on the frequency of query term in the document and in the whole collection. The algorithm next term is to find the most relevant terms for each other. The combination of these algorithms enhances the efficiency of the search for different information needs.

Keywords: *ranking, relevance, indexing, term, the search engine, the document, the token, algorithm.*