УДК 681.3

Т.Б. Гулка
Національний університет “Львівська політехніка”,
кафедра “Інформаційні системи та мережі”

МЕТОДИ АДАПТАЦІЇ ВЕБ-СТОРІНОК ДО ВИМОГ ПОШУКОВИХ СИСТЕМ

The review of features of information search of information resources to the Internet, the method of adaptation of structure of webs – pages under algorithms of search is offered.

Проведено огляд особливостей інформаційного пошуку інформаційних ресурсів в Інтернет, запропоновано метод адаптації структури веб-сторінок під алгоритми пошуку.

ВСТУП

Поява Інтернету й експонентне зростання обсягів його інформаційних ресурсів, безумовно, сильно стимулювали розвиток теорії інформаційного пошуку. Сьогодні більше ніж 75 % користувачів Інтернету використовують пошукові системи для доступу до інформації у глобальній мережі [1]. Специфіка Інтернету обумовила появу ряду нетривіальних задач як для розробників пошукових систем, так і для розробників веб-сайтів. Перші спрямовують свої зусилля на розробку нових алгоритмів якісного опрацювання запитів користувачів, другі – на вивчення цих алгоритмів та знаходження нових методів побудови веб-сайтів, адаптованих під алгоритми пошуку інформації. Основна проблема, що постає перед розробниками веб-сайтів, полягає в тому, що повні алгоритми роботи пошукових систем не публікуються через спроби нечесним шляхом досягнути високого рейтингу в результатах пошуку (наприклад, формування спеціальних сторінок для пошукових систем з білим текстом на білому фоні).

АНАЛІЗ ЛІТЕРАТУРНИХ ДЖЕРЕЛ

Пошукова система (ПС) – це програмний комплекс, що забезпечує доступ до множини слабоструктурованої інформації. У даному визначенні пошукової системи мається на увазі інформація різного роду, тобто текст, аудіо, відео, зображення тощо. Однак треба зазначити, що саме текстові дані ідеально підходять для опису повної функціональності пошукової системи, тому що алгоритми пошуку мультимедійної інформації насамперед ґрунтуються на алгоритмах пошуку тексту.

Основна задача пошукової системи – мінімізувати час, що витрачається користувачем на пошук релевантної запитові інформації. Релевантність – є одним із суб’єктивних і вельми заплутаних понять науки інформаційного пошуку. Найчастіше говорять про релевантність з погляду користувача, і у такому випадку “релевантна запитові інформація” і “потрібна користувачеві інформація” – одне і те ж поняття. У деяких обставинах релевантну інформацію визначають як всю інформацію з бази, що має відношення до запиту. Так, наприклад, якщо користувачеві потрібно довідатися все про конкретну фірму, то він зацікавлений у перегляді всіх документів, у яких є згадка про цю фірму. В інших випадках релевантна інформація – це тільки та інформація, що достатня для виконання визначеної задачі користувача, наприклад, пошуку

відповіді на конкретний запит. Якщо в останньому випадку в результатах пошуку буде багато надлишкових даних, тобто даних, що мають відношення до запиту, але не потрібні для виконання даної задачі, то вибірка потрібної/релевантної інформації займе в користувача додатковий час.

Отже, традиційно до пошукової системи застосовують дві основні характеристики: точність і повнота. Щораз, коли користувач формує до системи запит, тим самим ініціалізуючи пошук, усі документи в колекції пошукової системи поділяються на чотири частини, як це показано на рис 1.

Рис. 1. Розбиття простору документів при опрацюванні запиту за складом

Тоді точність (P) і повноту (F) можна визначити так

$$P = \frac{R \cap A}{A},\ F = \frac{R \cap A}{R}.$$

Точність результату пошуку показує, наскільки пошукова система здатна мінімізувати час, затрачений користувачем на пошук релевантної даному запитові інформації. Тоді як повнота результату пошуку характеризує інший аспект – наскільки добре система знаходить релевантну даному запитові інформацію. Можна підібрати оптимальний запит(и), коли кожен знайдений документ буде релевантним, і кожен релевантний документ буде знайдений. Крива “точність/повнота”‘ для оптимального запиту зображена на рис.1 ліворуч, однак для реальних запитів ця крива для більшості пошукових систем має вигляд, зображений на рис. 2 праворуч.

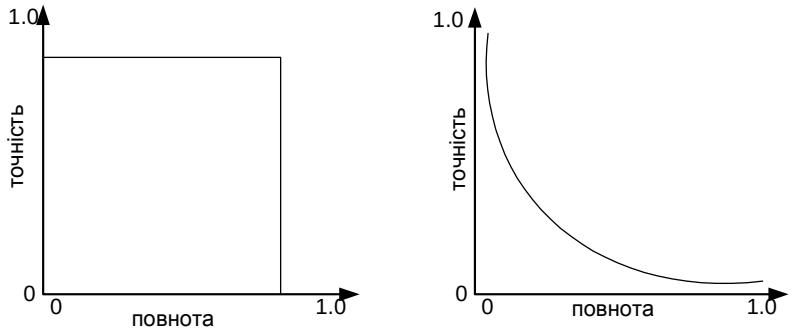

Рис. 2. Оптимальна і реальна залежність “точність/повнота”

Характеристики та параметри WWW можна розділити на характеристики WWW як набору даних та поведінки користувачів пошукових систем у WW.

ХАРАКТЕРИСТИКА ТА ПАРАМЕТРИ WWW ЯК НАБОРУ ДАНИХ

Методи пошуку, використовувані в класичних пошукових системах, розроблялися і тестувалися на відносно невеликих і доволі однорідних колекціях, таких, як бібліотечні каталоги або колекції газетних статей. WWW як набір даних має ряд важливих особливостей:

- Розмір – за останні роки було безліч спроб оцінити розмір Веб, і, хоча оцінки не цілком збігаються, усі вони виходять з того, що у Веб утримується більш мільярда сторінок. З огляду на те, що розмір середньостатистичної сторінки становить 5–10 Кб, то неважко підрахувати, що йдеться про терабайти.

Зазначимо, що ці оцінки стосуються тільки тієї "поверхневої" частини Веб, що не схована за пошуковими формами і доступ до якої не вимагає попередній реєстрації або авторизації. Іншу, "сховану" частину Веб пошукові системи зазвичай не розглядають, але ж до неї належить безліч реально опублікованих великих баз даних.

- *Динаміка розвитку*. Інформація в Інтернет змінюється дуже динамічно: інформаційні ресурси з'являються, щезають, змінюють своє місце розташування або зміст дуже часто. Так, щомісяця змінюється близько 40 % інформації, середній час життя половини сторінок у Веб не перевищує 10 днів. Зазначимо, що при цьому змінюється і використовувана лексика. У класичних пошукових системах тимчасові характеристики інформації практично не враховувалися.
- *Неструктурованість і надмірність*. Вважають, що Веб – це розподілений гіпертекст. Однак це не зовсім так. Гіпертекст зазвичай передбачає наявність концептуальної моделі, що накладає обмеження погодженості на дані і гіперзв'язки. У Веб це, звичайно, не так навіть для тих його частин, що знаходяться під єдиним адміністративним контролем. Близько 30 % інформації у Веб є точними або приблизними копіями інших документів.
- *Неконтрольована якість*. Відсутність редакторського контролю над інформацією, що публікується у Веб, обумовлює проблеми з якістю – інформація може бути некоректною (наприклад, уже застарілою), помилковою, погано сформульованою, з масою граматичних помилок. За деякими оцінками, одна помилка зустрічається в середньому в кожному двохсотому з часто вживаних слів або в кожному третьому іноземному прізвищі [1].

ОСОБЛИВОСТІ ПОВЕДІНКИ КОРИСТУВАЧІВ У ВЕБ

У Веб змінюється і поняття "типового користувача". Зазначимо такі відмінності:

- *"Погані" запити* – ніхто не навчає користувачів ІПС для Веб формулювати запити, і, як наслідок, значно менший відсоток користувачів використовує розширені можливості пошуку, такі як логічні вирази. Більш того, типові запити дуже короткі – більше ніж 60 % пошукових запитів у Веб складаються з 1–2 слів, що сильно відрізняється від 7–9 слів у класичних ПС.
- *Різнорідний контингент* – розмаїтість у знаннях, потребах і очікуванні користувачів дуже велика. Більшість же користувачів класичних ПС зазвичай має багато загальних рис.
- *Поводження* – користувач не готовий довго чекати результату і не готовий навіть шукати його в наданій системою вибірці. Так, 58 % користувачів не йдуть далі першої сторінки, що відображається на екрані, а 67 % не пробують модифікувати свій первісний запит.

Через перераховані вище особливості дуже важливою задачею в контексті Веб є таке впорядкування результатів пошуку, щоб першими виявилися ті результати, що є найбільш цікавими для користувача. Класичні підходи до рангування спираються на міру схожості тексту запиту і тексту документа – релевантності, але "розпливчасті" запити користувачів і величезна кількість документів значно знижують ефективність таких підходів у контексті Веб. Більше того, ніким не контрольована публікація у Веб дозволяє нечесно підвищувати рейтинг власної сторінки у результатах пошуку (наприклад, заповнюючи її "білим по білому" ключовими словами).

Тому більш перспективним є використання разом із релевантністю ще і міри важливості (корисності, популярності) Веб-сторінки при рангуванні результатів пошуку. Типовим прикладом такої метрики є індекс цитування, тобто кількість посилань на дану сторінку, що досить популярний у бібліометрії. Однак посилання у Веб сильно відрізняються від посилань у друкованій літературі, і ніщо не заважає авторові Веб-сторінки створити купу порожніх сторінок, що посилаються на дану. Тому необхідно враховувати важливість і сторінок, що посилаються.

Першим і найбільш відомим розширенням індексу цитування у Веб стала метрика PageRank, названа ім'ям одного із засновників пошукової системи Google (http://www.google.com) і реалізована в ній. Метрика PageRank рекурсивно визначає важливість сторінки p на основі інформації на сторінках q, що посилаються на сторінку q:

$$PageRank(p) = (1-d) + d \sum_{\forall q: q \to p} \frac{PageRank(q)}{links(q)}$$

де d – це деякий параметр (зазвичай порядку 0.85), а links(q) позначає кількість посилань, що виходять зі сторінки q.

ОПТИМІЗАЦІЯ ПАРАМЕТРІВ ВЕБ-СТОРІНОК

При рангуванні результатів пошуку в пошуковій системі дуже велике значення мають слова титульної фрази. А саме, якщо формулювання запиту збігається з титульною фразою або титульна фраза містить кілька слів запиту, то сторінка з таким титулом за інших рівних умов виявиться вище.

Титульна фраза (титул) – це текст, який міститься на сторінці в дескрипторі title, у коді сторінки він оформляється як <title>Титульна фраза</title>.

Варто зазначити суперечливий статус титульної фрази, що став причиною найбільш грубих помилок багатьох веб-майстрів. Титульна фраза не відображається на сторінці сайта в браузері. Точніше, відображається, але у заголовку вікна браузера, тобто в місці, куди зазвичай ніхто не дивиться. Саме тому настільки поширена помилка, коли веб-майстер ставить для всіх сторінок сайта однаковий титул. Переважно це буває сама назва сайта, що дуже коректно виглядає, коли сайт, наприклад, показується замовнику.

Але коли сайт подається на індексацію в пошукові системи, картина змінюється, тому що в результатах пошукових систем зміст тега <title> відображається як заголовок знайденої сторінки. Саме титульна фраза сторінки є найбільш яскравим елементом списку результатів пошуку за пошуковим запитом. Тут вона грає свою головну роль – заголовка, "обличчя" сторінки. І вибір користувача більш ніж наполовину визначається точністю, зрозумілістю і привабливістю титулу.

Отже, можна вивести перші правила підготовки сайта до індексації пошуковими системами:

1. Всі сторінки сайта повинні мати різні титули.

2. Титульна фраза кожної сторінки повинна бути точною і зрозумілою.

3. Зміст сторінки повинен відповідати титулу.

4. У титульній фразі сторінки повинні бути слова, які часто зустрічаються в пошукових запитах користувачів, відповіддю на які може бути дана сторінка.

5. Не варто використовувати титульні фрази довжиною понад 80 символів.

6. Якщо слова не містяться де-небудь у змісті (тексті) сторінки, то вони не повинні включатися в її тег Title.

Деже важливим є слова, які будуть тезі <title> розташовані і, в тому числі, порядок їхнього розташування. В тезі краще не використовувати одне слово більше двох разів і не допускати повторів більш ніж двох слів.

Найоптимальніший підхід при створенні сайта – це, в першу чергу, визначити такий набір ключових слів, що найкраще відбивають специфіку бізнесу, а потім вже скласти на основі даних слів конкретні тексти, що і складуть зміст сайта.

Для побудови титульної фрази можна використати закони Зіпфа (G.K. Zipf) на базі яких будуються майже всі алгоритми пошукових мащин.

Перший закон Зіпфа:

$$C = \frac{Fw * Rf}{Cw},$$

де C – ймовірність, з якою можна знайти слово в тексті (величина приблизно однакова для будь-якої мови), Fw – частота входження слова у текст, Rf – ранг частоти, Cw – кількість слів.

Якщо трохи перетворити формулу, то ця функція зводиться до типу y=k/x і її графік – рівностороння гіпербола. Отже, за першим законом Зіпфа: якщо найпоширеніше слово зустрічається в тексті, наприклад, 100 разів, то частота входження другого за популярністю слова з високою імовірність становитиме приблизно 50.

Другий закон Зіпфа “кількість – частота”

$$K = \frac{Cf}{Fw},$$

де K – деяка константа (однакова для будь-яких текстів однієї мови), Fw – частота входження слова у текст, Cf – кількість слів в Fw.

Дослідження показують, що якщо побудувати графік залежності частоти входження слова в текст від кількості слів, то найбільш значимі слова знаходяться в середній частині діаграми. Слова, які зустрічаються занадто часто, в основному виявляються сполучниками, прикметниками, в англійській мові – артиклями. Такі слова часто називають стоп-словами.

Від того, як буде виставлений діапазон значимих слів, залежить багато чого. Якщо його поставити широко – потрібні терміни “загубляться” серед допоміжних слів; встановити вузько діапазон – можна втратити суттєві терміни. Кожна пошукова система вирішує проблему по-своєму, керуючись загальним обсягом тексту, спеціальними словниками тощо.

Алгоритм пошуку ключових слів для побудови титульної фрази:

1. Відкриємо веб-сторінку та відкинемо всі HTML-теги.

2. Видаляємо з тексту стоп-слова.

3. Обчислюємо частоту входження кожного терміна. Регістр символів не враховуємо.

4. Будуємо рейтинг частоти.

5. Вибираємо діапазон частот. Він повинен знаходитися десь посередині. Не потрібно брати терміни, які зустрічаються занадто рідко чи занадто часто. Вибір діапазону суб’єктивний. Необхідно орієнтуватися на конкретний зміст тексту.

6. З обраного діапазону виписуємо терміни.

7. Модифікуємо тег <title>.

Наведемо результат роботи функції, що реалізовує алгоритм. Функція написана на PHP. Як тестовий текст взято HTML-сторінку із розділу допомоги веб-броузера Opera. Англійська мова вибрана через те, що в ній менше змін словоформ (менше змінюються закінчення через відсутність відмінків).

Приклад файла, що містить стоп-слова:

#a#the#of#it#in#to#if#for#is#on#do#and#

Результати роботи функції.

Таблиця 1

Результат роботи контрольного прикладу

слово	частота	слово	частота
opera	28	our	3
you	25	internet	3
as	12	used	3
be	9	an	3
that	8	support	3
this	8	web	3
most	7	including	3
browser	7	find	3
are	7	page	3
with	7	use	3
your	7	make	3
new	6	out	3
should	6	go	3
we	5	alphabetical	3
help	5	document	3
what	5	well	3
get	5	some	3
all	5	menu	3
or	5	dialog	3
the	4	software	3
will	4	complete	3
information	4	Інші слова з частотою меншою 3	
Menus	4		

Таблиця 2

Ранг частот

Частота	Ранг
1	318
2	54
3	21
4	4
5	6
6	2
7	5
8	2
9	1
12	1
25	1
28	1

Як видно з табл. 1, на результат вплинула велика кількість стоп-слів (відображені на сірому фоні), що не були включені у файл виключення. Крім того, багато стоп-слів не були видалені із тексту (наприклад, “the”). Це пов’язано з тим, що запропонована функція вважає, що стоп-слова в тексті відділені пробілами, а часто можуть стояти знаки пунктуації.

Окрім цього, на результат вплинув порівняно невеликий обсяг тестового тексту (приблизно одна друкована сторінка), проте це було зроблено свідомо, адже приблизно такий обсяг середньої веб-сторінки.

Враховуючи тематику сторінки, можна вибрати слова для титулу: opera, browser, help, internet, support, web, page, menu, software.

ВИСНОВОК

У результаті проведеного дослідження було розглянуто основні проблеми, що виникають при інформаційному пошуку в Інтернет та запропоновано один із способів адаптації існуючих веб-сторінок для збільшення їх рейтингу у пошукових системах. Розглянутий тестовий приклад довів ефективність запропонованого підходу. Надалі для покращання роботи функції доцільно:

- розробити ефективну технологію автоматизованого поповнення словника стоп-слів;
- удосконалити блок пошуку та видалення стоп-слів;
- розробити модуль аналізу структури слова для розпізнання однокореневих слів;
- розробити модуль автоматичної адаптації тегу <title>

Запропонований підхід також можна використовувати для модифкації інших частин веб-сторінки: мета-тегів, ключових слів, безпосередньо самого наповнення сторінки.

1. Буров Є.В., Пелещишин А.М. Оптимізація розміщення даних у Web-системах. // Інформаційні системи та мережі // Вісн. Держ. ун-ту "Львівська Політехніка". – 1998. – № 330. – С. 17–27. 2. Добрынин В.Ю., Некрестьянов И.С. Задача выбора тематических коллекций, релевантных запросу // Тр. Всерос. науч.-метод. конф. "Интернет и современное сообщество". – Санкт-Петербург, декабрь 1998. 3. Гринберг И., Ли Гарбер. Разработка новых технологий информационного поиска // Открытые системы. – 1999. – 10. 4. Степанов В.К. Русскоязычные поисковые механизмы в Интернет // ComputerWorld Россия. – 1997. – 11. 5. Галайко В.М. Розроблення інтелектуальних Web-систем // Вісн. Держ. ун-ту "Львівська політехніка". – 1998. – № 330. – С. 53–62. 6. Зайцев С.С. Описание и реализация протоколов сетей ЭВМ. – М.: Энергия, 1980. – 155 с.