УДК 004. 738 52

А.В. Вдовіченко
Житомирський інженерно-технологічний інститут, Україна

Інтелектуалізовані пошукові системи. Класифікація та порівняння

У статті розглянуті характеристики існуючих пошукових систем і запропонована їх класифікація відповідно до цих характеристик. Наведені порівняльні приклади для закордонних і вітчизняних пошукових машин і сформульовані висновки щодо необхідності їх удосконалення.

Вступ

Світова мережа Інтернет може бути представлена як гетерогенний інформаційний простір, що утворений з ресурсів різних типів та має довільну інформаційну структуру. Інформаційний простір змінюється безупинно, обсяг інформації у світі невпинно зростає, інформаційні структури стають більш складними, рівень хаосу в інформаційній системі постійно зростає. У світі, де інформаційні ресурси необмежені, головною проблемою та найважливішою цінністю стає не сама інформація, а саме пошук необхідної інформації та ранжування знайденої інформації відповідно до поставленого запиту [1-6]. Така постійно зростаюча практична потреба в пошуку інформації, зумовлена бурхливим розвитком інформаційних технологій та глобальних мереж, стимулює зацікавленість у пошукових системах і в сервісах, що вони реалізують.

Пошукові машини, як то: «crawlers», «spiders», «worms» – індексують мережу Інтернет на підставі алгоритмів штучного інтелекту та за заданими запитами користувачів [4]. Але, наприклад, у чому різниця (не схожість, а саме різниця) між запитами «Мені потрібен новий автомобіль» і «Хочу нову машину» для алгоритмів штучного інтелекту? Її немає. Сьогоднішні пошукові алгоритми, як правило, не розділяють, а поєднують синоніми. «Автомобіль» логічно прирівнюється до «автомашини», хоча соціологічні дослідження чітко виділяють фінансові й інші особливості груп (читай – «груп потенційних клієнтів»), що використовують ті чи інші слова та вирази. Таким чином, алгоритмізація, упорядкування мови знищує значеннєві відтінки. Зі збільшенням кількості проіндексованих сторінок, на жаль, часто знищується і сам зміст. Релевантність пошукової машини різко падає.

Каталожна організація інформації припускає, що укладач каталога добре обізнаний на *предметній області*, що охоплюється каталогом. «Предметна область» сьогоднішнього Інтернету – це досвід, накопичений людством упродовж своєю історії. До цього необхідно додати всю відображувану динамічну інформацію про сьогоднішній день і, можливо, дещо про завтрашній. Найбільш зручними для користувача є не занадто великі, спеціалізовані тематичні каталоги

з малою кількістю піддиректорій, і будь-які спроби створення глобального інформаційного каталога неминуче призводять до того, що безліч ефективних малих списків губляться в плутанині розділів і підрозділів. Таким чином, ускладнення каталожних систем знищує простоту користування ними. У результаті погіршується *релевантність* пошукової машини [5].

Будь-яку пошукову систему можна описати, виділивши певний набір характеристик. Спробуємо визначити деякі характеристики та визначити їх.

Характеристики пошукових систем

Характеристики, що залежать від структури документу

Більшість інформаційних ресурсів сьогоднішнього Інтернету представлені у вигляді статичних чи динамічних HTML, і пошукові системи повинні вміти індексувати ці документі, враховуючи їх внутрішню будову [2].

Підтримка фреймів. Пошукова система повинна вміти обробляти документи фреймової структури. Такі документи є дуже розповсюдженими, і втрата можливості їх індексування призведе до неможливості індексування багатьох документів, на які посилається фреймова структура.

Підтримка ImageMap. Тут приблизно та ж проблема, що і з фреймовими структурами серверів.

Частота появи посилань. Пошукові машини можуть визначити популярність документа за частотою посилань на нього з інших документів мережі, і, таким чином, вони визначають релевантність документа.

Стоп-слова. Деякі слова недоречно включати до індексу пошукової машини. Це, як правило, слова, що дуже часто віористовуються: чи у звичайній мові, чи в певній предметній області. Не включають такі слова для економії місця на носіях. Наприклад, Altavista ігнорує слово *web*.

Підтримка META-тегів. Теоретично, усі пошукові машини повинні враховувати метадані при індексації сторінок, однак на практиці не все це роблять.

Title. Цей параметр показує, як пошукові машини генерують заголовки посилань для користувача у відповідь на його запит.

Description. Цей параметр показує, як пошукові машини генерують опис посилань для користувача у відповідь на його запит.

Підтримка META-тегів. Усі пошукові машини при індексації сторінок повинні враховувати ці теги як засіб опису документа, однак на практиці не все це роблять.

Title. Показує, як пошукові машини генерують заголовки посилань для користувача у відповідь на його запит.

Description. Показує, як пошукові машини генерують опис посилань для користувача у відповідь на його запит.

Перевірка статусу URL. Визначає, чи можна перевірити, наскільки глибоко проіндексований сервер і чи є він взагалі в індексі пошукової машини.

Мовні характеристики

Пошукові системи можуть підтримувати одну чи більше мов документів для індексації та, відповідно, мов запитів користувача. Чим більше мов підтримує така система, тим вона більш розвинута, оскільки розширюється інформаційний простір для пошуку. Для якісного пошуку також важливо, які запити підтримує пошукова система, чи дозволяє вона використовувати повнотекстові та розширені запити.

Прямий запит. Прямий збіг фрази запиту й інформації в документі.

Запит «ЧИ»/«І». Документ повинен містити всі слова запиту («І»); документ повинен містити хоча б одне слово запиту («ЧИ»).

Повнотекстовий запит. Задається природною мовою у вигляді речення, побудованого за правилами цієї мови.

Розширений запит. Використання розширених виразів у запиті: використання метасимволів та регулярних виразів.

Генерація словоформ. Параметр визначає, чи спроможна пошукова система генерувати словоформи заданих слів при побудові запиту.

Технічні характеристики індексування

Зазначені (submitted) сторінки. Сторінки серверів набагато раніше з'являються в індексах систем, якщо їх прямо вказати (Add URL), незважаючи на те, що в ідеальному інформаційному просторі пошукова машина повинна проіндексувати будь-яку сторінку будь-якого сервера.

Незазначені (non-submitted) сторінки. Зазначивши хоча б одну сторінку сервера, в ідеалі пошукова машина обов'язково знайде та проіндексує всі інші сторінки за посиланням із зазначеної. Це, звичайно, потребує часу, і в реальному світі пошукові машини залишають індексування таких сторінок на майбутнє [6].

Контроль індексації. Визначає засоби, за допомогою яких відбувається керування пошуковою машиною. Для «розвинутих» пошукових систем такий контроль виконується розпорядженнями файла robots.txt. Деякі також підтримують контроль за допомогою META-тегів із самих документів, що індексуються.

Індексування захищених паролями директорій і серверів. Деякі пошукові машини можуть індексувати такі сервери, якщо їм указати Username та Password. Це може стати у нагоді користувачам, принаймні вони зможуть побачити, що є на сервері.

Глибина індексування. Визначає, скільки незазначених сторінок після зазначеної буде проіндексовано пошуковою системою. Більшість великих машин не мають обмежень щодо глибини індексування, але існують причини через які сторінки можуть бути не проіндексовані, як то: занадто акуратне використання фреймових структур без дублювання посилань у керуючому (frameset) файлі чи використання imagemap без дублювання їх звичайними посиланнями.

Видалення старих даних. Визначає дії, що виконуються при закритті сервера чи переміщенні його на іншу адресу. Можливі дві дії: просто видалити старий уміст і переписати файл robots.txt.

- *видалення вмісту:* проведення вилучення даних в індексі пошукової машини при спробі реіндексувати документ. Залежить від періоду відновлення даних;
- *robots.txt:* всі посилання на файли цього сервера будуть вилучені з індексу при звертанні пошукової машини до файла, якщо сервер весь закритий від індексації.

Перенапрямок (redirect). Деякі сайти перенаправляють відвідувачів з одного сервера на інший. Цей параметр указує, який URL буде зв'язаний із вашими документами. Це важливо, оскільки якщо пошукова машина не відпрацьовує перенапрямок, то можуть виникнути проблеми з неіснуючими файлами.

Період відновлення. Динаміка зміни документів в Інтернеті дуже велика й пошукові машини повинні індексувати всі без обліку дати. Однак посилання, видані у відповідь на запити користувачів, можуть бути одноденної давнини, а можуть бути й місячної давнини, а то й більше. Деякі пошукові машини відразу індексують одну

сторінку, а лише потім продовжують індексувати ще не проіндексовані сторінки, інші можуть частіше сканувати найбільш популярні сторінки мережі.

Дата індексування документа. Деякі пошукові машини показують дату, коли був проіндексований той чи інший документ. Це допомагає користувачу зрозуміти, якої «свіжості» посилання видає пошукова система.

Spam-штрафи. Підвищити рейтинг сайту можна завдяки багаторазовій вказівці себе через Add URL чи багаторазовому згадуванню того самого ключового слова тощо. Звичайно, пошукові системи намагаються уникати подібних дій і, навпаки, знижують рейтинг такого сайту.

Інтелектуальні характеристики

Crc32. Можна легко побачити, як при використанні пошукової системи вона на запит видає декілька посилань на один й той же документ, що не тільки знижує зручність використання системи, а й значно впливає на її загальну релевантність.

Звертання до інших систем. Деякі пошукові системи не мають власної бази індексів, а звертаються по інформацію до інших пошукових систем, передаваючи їм запит і отримуючи результат пошуку, а потім виконують ранжування отриманих даних.

Урахування популярності документа. Розвинуті пошукові машини при обрахуванні релевантності документа враховують його популярність: як часто на нього посилаються інші документи.

«Здатність до навчання»:

1. Якщо сервер оновлюється часто, то пошукова машина частіше буде його реіндексувати, якщо рідко – рідше.
2. Деякі пошукові системи здатні запам’ятовувати успішно виконані запити, щоб надалі виконувати їх більш ефективно.
3. Настроювання на роботу з окремим користувачем пошукової системи, створення персонального облікового запису з метою адаптації системи на стиль окремого користувача.

Вплив на алгоритм визначення релевантності. Пошукові машини обов'язково використовують розташування та частоту повторення ключових слів у документі. При цьому методи підвищення релевантності різні для кожної окремої машини.

Загальні характеристики

Тип пошукової машини. Повнотекстові пошукові машини індексують кожне слово на веб-сторінці, виключаючи *стоп-слова*. *Абстрактні* пошукові машини створюють деякий екстракт кожної сторінки. Для *веб-мастерів* повнотекстові машини корисніші, оскільки будь-яке слово, що зустрічається на веб-сторінці, підлягає аналізу при визначенні його релевантності до запитів користувачів. Однак може трапитися й таке, що для абстрактних пошукових машин сторінки проіндексовані краще, ніж для повнотекстових. Це може виходити з алгоритму екстрагування, наприклад, за частотою вживання на сторінці тих самих слів.

Ім'я пошукового робота. Ім'я пошукового робота – слово чи словосполучення, яким пошукова система відповідає на HTTP-запит.

Розмір пошукової машини – це кількість проіндексованих нею сторінок. Пошукові машини з великим розміром можуть проіндексувати майже всі сторінки сайту, машини середнього розміру проіндексують його лише частково, а при малому розмірі сторінки можуть узагалі не потрапити до каталогів пошукової машини.

Визначивши характеристики, згрупуємо їх, отримуючи таким чином класифікацію пошукових машин за даними характеристиками.

Дана класифікація відображена у табл. 1.

Таблиця 1

Характеристики пошукових машин

Загальні	Тип пошукової машини	
	Ім'я пошукового робота	
	Розмір	
Технічні характеристики індексування	Загальні	Час індексування вказаних сторінок
		Час індексування невказаних сторінок
		Контроль індексації
		Індексування захищених паролями директорій і серверів
		Глибина індексування
		Видалення старих даних
		Перенапрямок
		Використання META-тега для роботів
		Spam-штрафи
	Часові	Дата індексування документа
		Період відновлення
Структура документа	Підтримка фреймів	
	Підтримка ImageMap	
	Частота появи посилань	
	Стоп-слова	
	Підтримка META-тегів	
	Опис документа	
	Заголовок документа	
	Використання robots.txt	
	Перевірка статусу URL	
Інтелектуальні особливості	Crc32	
	Звертання до інших систем	
	Урахування популярності документа	
	Вплив на алгоритм визначення релевантності	
	Здатність до навчання	
Мова	Мова запиту	Прямий запит
		Запит «ЧИ»/«І»
		Повнотекстовий запит
		Розширений запит
		Генерація словоформ
	Мова документу	Одномовна
		Багатомовна

Наведемо опис існуючих інтелектуалізованих пошукових машин згідно з даною класифікацією. При цьому окремо виділимо закордонні та вітчизняні пошукові машини (табл. 2 і табл. 3).

Таблиця 2

Характеристики закордонних пошукових машин

	Altavista	Excite	HotBot	InfoSeek	Lycos	OpenText	WebCrawler
Тип	повнотекст	повнотекст	повнотекст	повнотекс	абстрактна	повнотекст	повнотекст
	1	2	3	4	5	6	7
Структура документа							
Підтримка фреймів	–	+	–	+	+	–	–
Підтримка ImageMap	+	–	–	+	+	–	+
Частота появи посилань	–	–	+	–	+	–	+
Стоп-слова	+	+	+	–	+	–	–
Підтримка META-тегів	+	–	+	+	+	–	–
Заголовок документа	Заголовок сторінки чи No Title	Заголовок сторінки чи Untitled	Заголовок сторінки чи URL	Заголовок сторінки чи перший рядок документа	Заголовок сторінки чи перший рядок документа	Перші 100 символів з документа	Заголовок сторінки чи URL
Опис документа	META-тег чи перші кілька рядків з документа	Форму-ється з найбільш релевант-них до запиту фраз документа	META-тег чи перші кілька рядків документа	META-тег чи перші 200 символів після тега <body>	META-тег чи екстракт із умісту сторінки	Перші 100 символів документа	Створю-ється з умісту; планується підтримка META-тегів у майбут-ньому
Використання META-тега для роботів	+	–	+	+	+	–	Тільки тег NOINDEX
Інтелектуальні особливості							
Crc32	–	–	–	–	–	–	–
Звертання до інших систем (*)	–	–	–	–	–	–	–
Урахування популярності документа	–	–	–	–	–	–	–
Вплив на алгоритм визначення релевант.	Немає	–	Ключові слова в метаданих	Немає	Немає	Немає	Частота появи посилань
Індексація з урахуванням частоти оновлення сервера	+	–	+	+	–	–	–
Запам'ятовування запитів	+	–	–	–	–	–	+

Продовження таблиці 2

	1	2	3	4	5	6	7
Обліковий запис користувача	–	–	–	–	–	–	–
Мова запиту та підтримка багатомовності							
Прямий запит	+	+	+	+	+	+	+
Запит «ЧИ»/«І»	+	+	+	+	+	+	+
Повнотекстовий запит	–	–	–	–	–	–	–
Розширений запит	+	–	–	–	–	–	–
Багатомовна	+	–	–	–	+	+	+
Технічні характеристики							
Час індексування вказаних сторінок	1 день	1 тиждень	3 тижня	1 місяць	1 місяць	2 – 4 тижні	2 – 4 тижні
Час індексування невказаних сторінок	1 – 3 місяці	3 тижні	3 тижні	1 місяць	1 місяць	2 – 4 тижні	2 – 4 тижні
Контроль індексування	robots.txt	robots.txt і метадані	robots.txt і метадані	robots.txt	robots.txt	robots.txt	robots.txt і метадані
Індексування захищених паролями директорій і серверів	–	+	–	+	+	–	–
Глибина індексування	Необмеж.	Необмеж.	Необмеж.	Необмеж.	Необмеж.	–	Обмежена популярністю серверу
Видалення старих даних	Видалити вміст і вказати нову адресу	Видалити вміст, переписати robots.txt	Переписати robots.txt	Видалити вміст і вказати нову адресу чи переписати robots.txt	–	–	–
Перевірка статусу URL	+	–	–	–	+	–	–
Перенапрямок	+	+	–	–	–	–	+
Дата індексування	Явне використ.	–	Неявне використ.	–	–	–	–
Період поновлення	від 1 дня до 3 місяців	1 – 3 тижня	не пізніше 3 тижнів	від хвилин до місяця	Щомісяч. Відновлення	1 – 4 тижня	щотижневе відновлення

Продовження таблиці 2

	1	2	3	4	5	6	7
Spam-штрафи	+	+	+	+	+	+	+
Використання robots.txt	+	–	+	+	+	–	+
Ім'я пошукового роботу	Scooter	Architext Spider	Slurp the Web Hound	Side winder	T-rex	–	Spidey
Розмір (ст.)	30 млн.	55 млн.	54 млн.	20 – 50 млн.	20 – 25 млн.	5 млн.	2 млн.

Таблиця 3

Характеристики вітчизняних пошукових машин

	Russian Express	TELA поиск	Rambler	Яndex	Апорт Поиск
Тип	повнотекст	повнотекст	повнотекст	повнотекст	абстрактна
	1	2	3	4	5
Структура документа					
Підтримка фреймів	+	+	+	+	+
Підтримка ImageMap	+	+	+	+	+
Частота появи посилань	–	–	+	–	+
Стоп-слова	–	–	+	+	+
Підтримка META-тегів	+	+	+	+	+
Заголовок документа	Поки URL	title	title чи URL і відносна міра релевантності	title і URL	Title
Опис документа	META-тег Description і частина тексту документа	Перші рядки документа	Перші 512 байт документа крім meta, javascript, images...	Видаються перші 1024 байт тексту, міра релевантності, дата створення й обсяг документа	Пропозиції, що містять слова запиту (1, 3 чи до 10)
Використання META-тегу для роботів	+	–	+	+	+
Інтелектуальні особливості					
Crc32	–	–	–	–	–
Звертання до інших систем (*)	–	–	–	–	–
Урахування популярності документа	–	–	–	+	–
Вплив на алгоритм визначення релевантності	–	–	–	–	–
Індексація з урахуванням частоти оновлення сервера	+	–	+	+	–

Продовження таблиці 3

	1	2	3	4	5
Запам'ятовування запитів	–	–	+	+	–
Обліковий запис користувача	–	–	–	–	–
Мова запиту та підтримка багатомовності					
Прямий запит	+	+	+	+	+
Запит «ЧИ»/«І»	+	+	+	+	+
Повнотекстовий запит	–	–	–	–	–
Розширений запит	–	–	+	+	–
Багатомовна	–	–	+	+	+
Технічні характеристики					
Час індексування вказаних сторінок	20 днів	–	7 – 14 днів	1 – 2 дня	1 – 15 днів
Час індексування невказаних сторінок	20 днів	–	до 3 місяців	залежно від популярності документів	лімітується швидкістю відновлення індексу
Контроль індексування	–	Явно – ні, побічно – вказавши як критерій URL	+	+	+
Індексування захищених паролями директорій і серверів	–	+	–	+	+
Глибина індексування	5000 документів на глибину 150	20 документів	Необмежена	Необмежена	Необмежена
Перевірка статусу URL	+	–	–	–	+
Перенапрямок	+	+	–	–	–
Дата індексування	Ні, у проекті – так	+	Так, при розширеній видачі результатів	+	+
Період поновлення	20 днів	3 – 4 тижні	1 раз на тиждень	Перманентно	раз на добу (від 10 до 40 тисяч документів)
Spam-штрафи	+	+	+	+	+
Використання robots.txt	+	+	+	+	+
Ім'я пошукового робота	www.search.ru	–	StackRambler/1.2	YandexWeb	Aport
Розмір (стор.)	500000	500000	2500000	2000000	2600000

Висновки

За даною класифікацією та за наведеними характеристиками реальних пошукових систем, які функціонують сьогодні, можна зробити висновки щодо невеликої ефективності їх функціонування. Більшість існуючих сьогодні

пошукових систем орієнтовані, перш за все, на так званий «масовий» пошук: увага приділяється кількості проіндексованих документів та швидкості проведення пошуку. Пошукові системи не охоплюють усіх інформаційних ресурсів і майже не здійснюють контролю вірогідності отриманої інформації. Можна стверджувати, що *якість* пошуку дуже низька: характеристики, що визначають релевантність системи (інтелектуальні характеристики, мовні характеристики), показують, що існуючі пошукові системи потребують удосконалення. Пошукові системи потребують інтелектуалізації для досягнення більшої релевантності: розширення інтелектуальних і мовних характеристик, інтелектуалізації обробки текстової інформації та покращання методів ранжування знайдених документів.

Література

1. Талантов М. Профессиональный поиск в Інтернет. Полнота, достоверность, скорость // КомпьютерПресс. – 1999. – № 7.
2. Солтон Дж. Динамические библиотечно-информационные системы. – М.: Мир, 1979.
3. Pinkerton Brain Finding What People Want: Experiences with the WebCrawler // http://info.webcrawler.com/bp/WWW94.html
4. Тихонов В. Поисковые системы в сети Інтернет // atomzone.hypermart.net
5. Pinto Francisco, Baptistay Claudio, Ryanz Nick. Using Semantic Searching for Web Portal Interoperability
6. Аликберов А.А. Подборка статей и технической информации по работе и характеристикам поисковых машин // http://www.citforum.ru/internet/search/ips.shtml
7. Alta Vista http://www.altavista.digital.com/ Digital Equipment Corporation, 1996.

The features of existing searching systems and their classification are considered. The comparative examples of foreign and home searching systems are presented. The conclusions of their needs of improvement are presented.

Статья поступила 01.07.02.