kasia szymczak

Szukanie pod lupą

Wyszukiwarki to najbardziej popularne serwisy internetowe. Wszyscy z nich korzystamy, ale mało kto wie, jak wśród miliardów dokumentów potrafią znaleźć te, które nas interesują.

Internet jest największym zbiorowiskiem informacji, stworzonym przez człowieka. Znaleźć można tu dokładnie wszystko, począwszy od przepisów na rosół z kury, a na nitroglicerynę skończywszy. Jednak taki ogrom informacji niesie ze sobą bardzo poważny problem – jak odszukać to, co nas interesuje?

Istnieją dwa rozwiązania tego problemu: katalogi oraz wyszukiwarki. Pierwsze są budowane przez duże sztaby osób, które przeglądają i klasyfikują ciekawe strony. Najstarszym i najsłynniejszym przykładem takiego serwisu jest Yahoo!. Katalogi przechowują informacje o ograniczonej liczbie dokumentów, jednak są to zazwyczaj najlepsze strony, jakie można znaleźć na dany temat na Internecie.

Twórcy wyszukiwarek podeszli do tego zagadnienia z nieco innej strony. Serwisy te automatycznie budują bazy danych zawierające informacje o olbrzymiej liczbie stron internetowych. Dokumenty te nie są jednak poukładane tematycznie, ponieważ komputer nie jest w stanie ocenić, o czym traktuje dana strona. Są natomiast poindeksowane według wyrazów, które zawierają. Pozwala to wyszukać te strony, o które użytkownik pytał.

Wyszukiwarki dzielą się również na dwie grupy: globalne i lokalne. Zadaniem pierwszych z nich jest indeksowanie wszystkich stron WWW – zadając w nich pytania, otrzymamy odnośniki do dokumentów znajdujących się na całym świecie. Przykładami takich serwisów są Infoseek oraz AltaVista. Natomiast wyszukiwarki lokalne starają się jak najdokładniej przeskanować strony znajdujące się w jednym kraju. Jedną z takich usług jest obecny na stronie głównej CHIP-a Online NET-oskop, który przeszukuje wyłącznie polski Internet.

Odrobina historii

World Wide Web powstała w 1991 r. W tym roku został opublikowany protokół HTTP (Hypertext Transfer Protocol), za pomocą którego przeglądarki komunikują się z serwerami. Pierwszy serwer WWW znajdował się pod adresem *http://info.cern.ch/*. W 1993 r. powstała graficzna przeglądarka WWW o nazwie „Mosaic".

Natomiast pierwsza prawdziwa wyszukiwarka, o nazwie *WebCrawler*, powstała w roku 1994. Różniła się od poprzednich tego typu projektów tym, że indeksowała całe strony, a nie tylko ich tytuły. WebCrawler wyszukiwał dokumenty znajdujące się na ok. 6 000 serwerów WWW. W tym samym roku powstał najpopularniejszy katalog Yahoo!.

Kilka miesięcy później wystartowała wyszukiwarka *Lycos* (*http://www.lycos.com/*), która szybko wyprzedziła WebCrawlera, gdyż była w stanie nadążyć za zawrotnym tempem rozwoju WWW. W lipcu 1994 r. baza danych Lycosa zawierała informacje o 54 000 dokumentów. W sierpniu liczba ta wzrosła już do 394 000, w styczniu 1995 do 1,5 miliona, a pod koniec 1996 – 60 milionów.

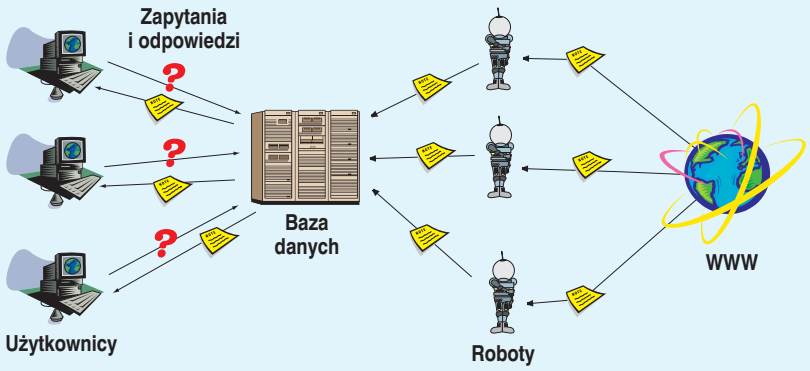
W 1995 r. powstała kolejna popularna usługa: *Infoseek* (*http://www.infoseek.com/*). Wyszukiwarka ta do dziś wyróżnia się przyjaznością dla użytkownika. W tym samym roku na rynek weszła również *AltaVista* (*http://www.altavista.com/*). Chociaż pojawiła się dość późno, jako pierwsza była wyposażona w wiele opcji, które dziś uważa się za standardowe. Wprowadzono w niej możliwość zadawania pytań w trybie logicznym, przy użyciu operatorów AND, OR oraz NOT, pozwalała też zadawać pytania w języku naturalnym. Była również najszybszą ze wszystkich wyszukiarek.

Najnowszym graczem na rynku wyszukiwarek jest *Google* (*http://www.google.com/*). W odróżnieniu od swoich konkurentów, którzy wykorzystują kilka potężnych komputerów, Google działa na tysiącach tanich pecetów złączonych w tzw. klaster. Dzięki temu ma w swojej bazie informacje o największej liczbie stron: ok. 1 300 000 000. Po raz kolejny okazało się, jak trudno skonstruować wyszukiwarkę, która potrafi nadążyć za przyrostem stron w Internecie.

Podróż do wnętrza

Wyszukiwarka jest po prostu bazą danych przechowującą informacje o zawartości zindeksowanych dokumentów. Wygląda to mniej więcej tak, że dla każdego słowa wyszukiwarka zapisuje adresy stron, na których występuje dany wyraz. W rzeczywistości informacji jest więcej – wyszukiwarki starają się ocenić, w jakim stopniu dane słowo opisuje treść strony, zapamiętywana jest więc miara ważności. Kiedy użytkownik korzysta z wyszukiwarki, program stara się znaleźć te stro-

Schemat działania wyszukiwarki



Internet jest przeszukiwany przez roboty. Każdy z nich pobiera z Sieci dokumenty, których treść jest indeksowana i zapisywana w bazie danych wyszukiwarki. Ta po otrzymaniu zapytania tworzy listę dokumentów i przesyła ją do użytkownika.

ny, na których znajdują się szukane słowa. Na początku listy dokumentów są te, które są najlepiej opisywane przez wyrazy z zapytania.

Pozostaje pytanie, skąd biorą się strony w bazie danych wyszukiwarki? Wypełnianiem bazy danych zajmują się roboty. Są to specjalne programy, zwane również pająkami, gdyż „łażą" po Sieci. Odczytują one odwiedzaną stronę WWW, dodają jej zawartość do bazy danych wyszukiwarki, po czym wyławiają z niej wszystkie odnośniki do innych dokumentów. Następnie odwiedzają świeżo znalezione strony i tak w kółko. Co jakiś czas w celu uaktualniena informacji odczytują też dokumenty, które już były odwiedzone. Wyszukiwarki pozwalają również użytkownikom wskazać adresy nowych stron, co sprawia, że zostają one natychmiast dodane do bazy. W tym celu każdy serwis ma formularz o nazwie „Dodaj stronę", gdzie znajduje się okienko do wpisywania adresów.

Taki mechanizm sprawia, że aby strona mogła znaleźć się w bazie wyszukiwarki, muszą kierować do niej odsyłacze z zewnątrz (po których trafią do niej roboty) albo ktoś musi podać jej adres wprost. W przeciwnym wypadku nie znajdzie się w ona odpowiedziach na żadne zadane pytanie.

Ziarno a plewy

Jednym z najtrudniejszych zadań wyszukiwarki jest zwrócenie stron w takiej kolejności, aby te najciekawsze dla użytkownika znalazły się na samej górze listy. Rozpoznawanie „związku" strony z danym pytaniem jest jednym z najtrudniejszych zagadnień w dziedzinie wyszukiwania informacji i różne wyszukiwarki stosują tu różne kryteria – najczęściej spotykane w literaturze naukowej nazywa się „miarą tf*idf". Zgodnie z tą zasadą wyraz dobrze opisuje dokument, jeśli pojawia się bardzo często na tej stronie i bardzo rzadko poza nią. Tak więc słowo „ale", które znajduje się na większości polskich stron, nie jest uważane za ważne, natomiast „Spectrum" dobrze charakteryzuje strony o komputerze Spectrum, gdyż rzadko występuje poza nimi.

Większość wyszukiwarek wzbogaca swój algorytm sortowania wyników dodatkową informacją, mającą związek z budową strony. Jeśli szukane słowo znajduje się w jej tytule, można sądzić, że opisuje ją lepiej, niż gdyby znajdowało się w jej treści. Wiele wyszukiwarek również wychodzi z założenia, że popularne strony są lepszymi odpowiedziami niż nieznane. Tak więc na liście znalezionych dokumentów wyżej pojawiają się adresy dokumentów, do których prowadzi wiele odnośników z innych stron w Sieci.

Tym, co również utrudnia życie projektantom wyszukiwarek internetowych, jest fakt, że użytkownicy rzadko zadają swoje pytania tak, aby dobrze opisywały interesujące ich strony. Bardzo dużo energii wkłada się więc w tworzenie rozmaitych systemów, które próbują odgadnąć, co tak właściwie użytkownik ma na myśli, zadając pytanie. Głównie robi się to na podstawie analizy poprzednich zapytań użytkownika i sprawdzania, które odnośniki z odpowiedzi użytkownik klika.

Pomoc dla robota

Autorzy stron WWW nie muszą współpracować z wyszukiwarkami. Jeśli jednak tego nie zrobią, jest mała szansa, że ktokolwiek te strony znajdzie. Główny mechanizm współpracy polega na umieszczaniu w nagłówku strony informacji, które wyszukiwarki użyją, aby ułatwić wyszukiwanie. Najważniejszy jest tu znacznik META o nazwie „description". Jego zawartość jest traktowana przez wyszukiwarki jako streszczenie strony. Jeśli dokument zawierający taki znacznik META pojawi się na liście wyników wyszukiwania, to pod jej tytułem będzie widniało właśnie to streszczenie. Chodzi oczywiście o to, aby pomóc szukającemu ocenić, czy znaleziona strona zawiera informacje, których szukał. Przykładowy opis wygląda następująco:

```
<meta name='description'
content='Strona domowa Gargamela.'>
```

Innym znacznikiem META, stworzonym na potrzeby wyszukiwarek, jest „keywords". Zawiera on słowa skojarzone z tematyką strony. Niestety – autorzy stron WWW rzadko wpisują tam sensowne wyrazy, więc część wyszukiwarek nie przywiązuje do tego znacznika większej wagi. Podajemy jednak, jak powinien on wyglądać:

```
<meta name='description' content=
'Gargamel, smerfy,papa smerf,
przepis na smerfa'>
```

Aby wyszukiwarka mogła dobrze zaprezentować stronę wśród swoich wyników, należy zadbać, aby tytuł strony oraz znacznik ME- w 170

Jak wyszukiwarka sprawdza, czy dwa dokumenty mają podobną treść

Na podstawie zawartości wszystkich stron WWW znajdujących się w Internecie jest tworzona tabela widoczna obok. Zawiera ona informacje, jak często dane słowa występują w różnych dokumentach.

P_{ij}, które określa, jak bardzo dane słowo jest ważne dla dokumentu P_j, obliczana jest wzorem $P_{ij} = tf_{ij} * id_{fi}$, gdzie:

Tf (term frequency) – współczynnik tym większy, im częściej słowo S_i występuje w dokumencie D_j;

Idf (inverted term frequency) – współczynnik ma tym większą wartość, im rzadziej słowo S_i pojawia się we wszystkich dokumentach Sieci.

Dokumenty

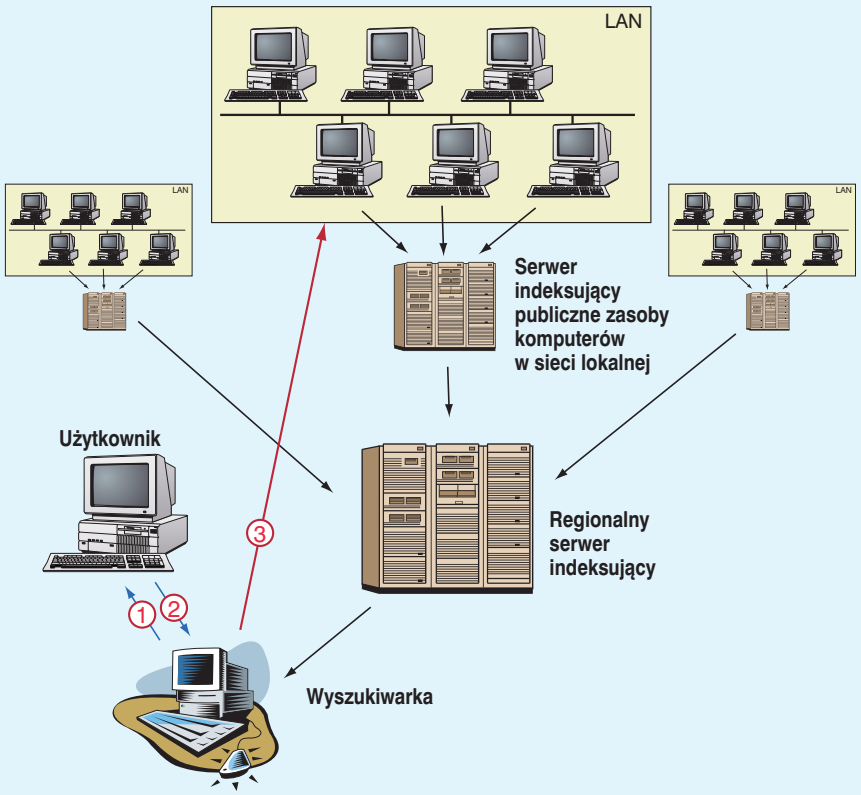
Słowa	D_1	D_2	.....	D_n
S_1	P_{11}	P_{12}	P_{13}	.....
S_2	P_{21}	P_{21}	.....	
.	P_{31}	.....		
.	.			
.	.			
.	.			
S_n	.			

Jeżeli wyrażenie:

$$\sum_{=}$$

jest duże, to dokumenty Dm i Dn mają podobną treść.

Wyszukiwarka przyszłości



Wyszukiwanie informacji w jednym miejscu jest niepraktyczne, gdyż pojedyncze komputery nie nadążają za rozwojem Internetu. W przyszłości każda sieć lokalna będzie indeksowała własne zasoby. Centralny komputer „przypyta" serwery lokalne albo regionalne, złoży ich odpowiedzi w jedną całość i prześle użytkownikowi.

CHIP

TA „description" dobrze opisywały stronę. Należy również oczywiście nakierować roboty wyszukiwarek na swoją stronę, by trafiła do ich baz danych.

Ostatni znacznik META, o którym wspomnimy, ma zadanie odwrotne do poprzednich – jest poleceniem, żeby roboty nie skanowały danej strony. Jest to potrzebne w wypadku dokumentów zawierających informacje, które są na tyle osobiste, że nie powinny być łatwe do znalezienia za pomocą wyszukiwarki. Znacznik wygląda następująco:

```
<meta name='robots'
value='noindex'>
```

Sztuka zadawania pytań

Jak już mówiliśmy, wyszukiwarka szuka stron związanych z zapytaniem na podstawie zawartych w niej słów. Skuteczne wyszukiwanie informacji polega więc na odpowiednim wyborze wyrazów, które użyjemy w zapytaniu. Należy unikać pojęć ogólnych albo takich, które mają kilka znaczeń. Jeżeli szukamy dokumentów na temat historii wyszukiwarek, wbrew pozorom zapytanie „historia wyszukiwarek" nie jest najlepsze, gdyż słowo „wyszukiwarka" jest zbyt ogólne. Można jednak skorzystać z informacji, że WebCrawler jest jedną z najstarszych usług tego typu, więc jej nazwa znajdzie się na każdej stronie o tej tematyce. Dużo lepsze wyniki da więc zapytanie „historia webcrawler". Oprócz wyboru dokładniejszych słów zapytania można zawężać, dodając słowa. Tak więc zapytanie „historia wyszukiwarek" zwróci gorsze wyniki niż np. „historia wyszukiwarek internetowych". W zawężaniu zapytań bardzo przydaje się też wyszukiwanie całych zwrotów. Jeśli otoczymy nasze zapytanie cudzysłowem, zwrócone zostaną tylko te strony, w których dane słowa występują jedno po drugim.

Większość wyszukiwarek na zapytanie użytkownika zwraca wszystkie te strony, które zawierają choć jedno słowo z zapytania. Algorytm sortujący stara się, aby dokumenty zawierające większą liczbę słów, o które pytaliśmy, były na górze. Czasami jednak chcemy bardziej dokładnie kontrolować, jakie adresy zostaną zwrócone. Wszystkie popularne wyszukiwarki mają odnośniki do specjalnego formularza, gdzie można zadawać pytania w trybie logicznym (advanced query). Wskazujemy wtedy dokładnie, które słowa powinny występować na szukanych stronach, a nawet jakie nie powinny na nich widnieć.

W tym celu korzystamy z operatorów logicznych AND, OR oraz NOT. Przykładowo, zapytanie „przepis AND rosół" wyszuka tylko te strony, które zawierają oba podane słowa. Zapytanie „przepis AND (rosół OR żurek)" zwróci strony, które zawierają słowo „przepis" i jedno ze słów „rosół" oraz „żurek". Natomiast zapytanie „przepis AND rosół AND NOT żurek" zwróci strony zawierające „przepis" oraz „rosół", ale nie zawierających słowa „żurek".

Wyjątkiem od tej reguły jest Infoseek, który pozwala zadawać logiczne zapytania w zwykłym formularzu. Słowa, które muszą się znajdować w zwróconych dokumentach, należy poprzedzić znakiem „+", a te, których nie powinno tam być – znakiem „-".

Jeśli na nasze zapytanie nie dostaniemy żadnych dokumentów, może to oznaczać jedną z dwóch rzeczy. Albo – co jest bardzo wątpliwe – w Internecie nie ma dokumentów na szukany przez nas temat, albo popełniliśmy literówkę, wpisując zapytanie do wyszukiwarki. W takich sytuacjach należy bardzo starannie sprawdzić pisownię.

Co dalej?

Wyszukiwarki internetowe są ciekawą i bardzo szybko rozwijającą się dziedziną. Największym problemem jest nadążanie za rozwojem Sieci. Okazuje się, że tradycyjne technologie i algorytmy sobie z tym nie radzą i trzeba wymyślać nowe sposoby przechowywania i przeszukiwania baz danych. Wyszukiwarka zawierająca najwięcej stron – Google – „przeskoczyła" konkurencję, rozpraszając swoją bazę danych na tysiącach niewielkich komputerów, niewiele szybszych od typowych komputerów biurowych.

Pomysł ten można podciągnąć o jeden szczebel wyżej – informacja może teoretycznie być wyszukiwana w miejscu, w którym się znajduje (patrz rysunek obok). Za wyszukiwanie odpowiedzialny byłby jeden serwer indeksujący całą sieć lokalną. Taki rozproszony system wyszukiwawczy wymagałby oczywiście opracowania protokołów, które pozwalałyby wyszukiwarkom komunikować się z hierarchiczną strukturą indeksujących serwerów. Rozmiar Internetu powoli sprawia, że utrzymywanie centralnej bazy danych w jednym miejscu staje się nieefektywne.

Jacek Surażski

INFO

Grupy dyskusyjne
Uwagi i komentarze do artykułu:
news://news.vogel.pl/chip.artykuly
Pytania techniczne:
news://news.vogel.pl/chip.internet

Internet
Opis wyszukiwarki FAST
http://www.fast.no/fast.php3?d=technology&c=fastsrch&h=2
Htdig
http://htdig.sourceforge.net/
Moduł w Perlu do tworzenia robotów
http://search.cpan.org/search?module=WWW::Robot