STUDIA INFORMATICA 2001
Volume 22 Number 3 (45)

Paweł KASPROWSKI
Politechnika Śląska, Instytut Informatyki

BUDOWA I SPOSOBY DZIAŁANIA WYSZUKIWAREK INTERNETOWYCH

Streszczenie. W pracy omówiono budowę wyszukiwarek internetowych, przeprowadzono analizę problemów występujących przy ich projektowaniu i przedstawiono możliwe rozwiązania z odnośnikami literaturowymi. W pierwszej części pracy wyjaśniono, na czym polega specyfika baz danych przechowywanych przez wyszukiwarki. W kolejnych omówiono trzy podstawowe elementy klasycznej wyszukiwarki internetowej odpowiedzialne za zbieranie, przechowywanie i udostępnianie danych.

ARCHITECTURE AND FUNCTIONING OF SEARCH ENGINES

Summary. This work describes how search engines work. Analysis of developing problems and possible solutions was presented with related literature. First section of this paper explains specific problems of text databases used in search engines. Next sections are introducing three elements of search engine architecture responsible for collecting, indexing and presenting text data. There are examples of specific problems

1. Wstęp

Problem opracowania sposobów pobierania, przechowywania i wymiany informacji jest podstawowym problemem informatyki. W dzisiejszych czasach systemy informatyczne „cierpią" z powodu nadmiaru informacji. Wobec ogromnego przeładowania danymi, problemem najważniejszym staje się opracowanie sposobów przetworzenia informacji i prezentacji jej w sposób, jakiego oczekuje użytkownik. Istnieją nowoczesne i bardzo wydajne narzędzia pozwalające na tworzenie bardzo dużych baz danych. Jednak wciąż brak narzędzi pozwalających na efektywne korzystanie z takich baz.

Klasycznym przykładem bazy przeładowanej trudno dostępną informacją jest wyszukiwarka internetowa (ang. search engine). Jest to element sieci, z którym spotyka się praktycznie każdy użytkownik Internetu. Każdy często napotykał z pewnością problemy ze znalezieniem pewnej specyficznej informacji. Nie zawsze możliwe jest określenie kryteriów wyszukiwania w sposób, który pozwoli na uzyskanie z wyszukiwarki konkretnej i użytecznej informacji.

W pracy omówiono budowę wyszukiwarek internetowych, przeprowadzono analizę problemów występujących przy ich projektowaniu i przedstawiono możliwe ich rozwiązania z odnośnikami literaturowymi. W pierwszej części pracy wyjaśniono, na czym polega specyfika baz danych przechowywanych przez wyszukiwarki. W kolejnej omówiono trzy podstawowe elementy klasycznej wyszukiwarki internetowej odpowiedzialne za zbieranie, przechowywanie i udostępnianie danych. W następnych zostały zaprezentowane problemy z oceną wyszukiwarek oraz przyszłość rozwoju tego typu systemów. Ostatnia część podsumowuje pracę.

2. Klasyczne bazy danych a bazy tekstowe

Przez bazę tekstową rozumie się w tej pracy bazę przechowującą informacje na temat zawartości dokumentów tekstowych. W szczególności przez dokument tekstowy w kontekście wyszukiwarek internetowych rozumie się stronę WWW zapisaną w języku HTML. Duża baza tekstowa jest szczególnym rodzajem bazy danych niosącym wiele problemów nieistotnych lub w ogóle nie występujących w klasycznych, relacyjnych bazach danych.

Przy uzyskiwaniu informacji z dużych baz danych napotyka się dwa podstawowe problemy: szybkość odpowiedzi i prawidłowość wyniku. Szybkość odpowiedzi poprawiać można za pomocą różnorodnych technik. Najprostszą i znaną od dawna techniką jest indeksacja bazy. Dzięki niej elementy w bazie mogą być znalezione relatywnie szybko. Przy bardzo dużych bazach danych efektywna staje się metoda wstępnej agregacji danych. Stosuje się ją w coraz popularniejszych hurtowniach danych. Polega ona, w dużym uproszczeniu, na tworzeniu dodatkowych gotowych informacji, powstałych w wyniku przetworzenia pewnej ilości informacji atomowych. Na przykład ze zbioru informacji na temat wysokości obrotów w sklepach na terenie miasta możemy wyliczyć i zapamiętać sumaryczny oraz średni obrót wszystkich sklepów w mieście. Taka informacja raz umieszczona w bazie może być wykorzystywana bez potrzeby kolejnego czasochłonnego przeliczenia wszystkich jej rekordów składowych.

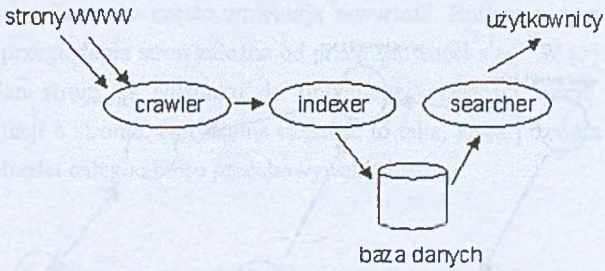
Jak łatwo zauważyć, hurtownie danych sprawdzają się tylko dla stabilnych danych liczbowych. Dla danych szybkozmiennych oraz nie podlegających łatwym agregacjom danych tekstowych, techniki używane w hurtowniach danych nie zdają zupełnie egzaminu. Do zwiększenia szybkości odpowiedzi w takich przypadkach stosuje się zwykle rozpraszanie przetwarzania w połączeniu ze zwiększaniem zasobów maszyn przetwarzających. Ta metoda sprawdza się również w dużych bazach tekstowych.

W klasycznej bazie danych nie ma problemu prawidłowości wyniku. Użytkownik definiuje kryteria i na ich podstawie otrzymuje odpowiedź lub zbiór odpowiedzi je spełniających. W bazie tekstowej bardzo trudno jest zdefiniować prawidłowe kryteria. Zwykle zbiór uzyskanych odpowiedzi prawidłowych jest tak duży, że najistotniejsze staje się oszacowanie, która z nich jest najwłaściwsza – najbardziej użyteczna z punktu widzenia użytkownika.

Istnieje wiele metod pozwalających przybliżyć się do wyniku najbardziej prawidłowego – niektóre z nich zostaną pokrótce przedstawione w niniejszej pracy.

3. Budowa wyszukiwarki internetowej

Wyszukiwarka internetowa to system składający się z kilku, współdziałających ze sobą, elementów. Każdy z tych elementów można realizować za pomocą różnorodnych technik i każdy niesie ze sobą pewne określone problemy informatyczne.



Rys. 1. Architektura wyszukiwarki internetowej
Fig. 1. Search engine archiecture

W kolejnych sekcjach omówiono trzy podstawowe elementy klasycznej wyszukiwarki internetowej odpowiedzialne za:

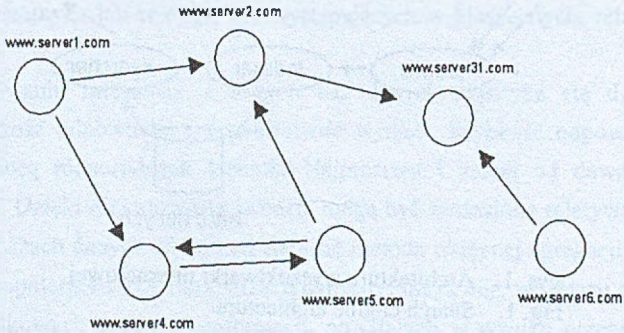
- zbieranie danych (crawler),
- zapisywanie danych (indexer),
- udostępnianie danych (searcher).

3.1. Zbieranie danych (crawler)

Zasoby internetowe wciąż rosną i podlegają modyfikacjom. Ilość informacji jest tak ogromna, że do lamusa odchodzą wyszukiwarki bazujące na ręcznym gromadzeniu danych przez pracowników firmy, tak jak robiono to u prapoczątków wyszukiwarki Yahoo. Sposób taki jest po prostu zbyt wolny – choć z pewnością pozwala na znacznie lepszą segregację danych, co z kolei zapewnia większą użyteczność informacji zawartych w bazie. Jednak nawet duża grupa ludzi nie jest w stanie stworzyć bazy o wielkości porównywalnej z bazą tworzoną automatycznie za pomocą komputerów. Poza tym oczywiście zatrudnienie „armii przeglądaczy" znacznie zwiększa koszty utrzymania wyszukiwarki.

Pobieraniem danych do baz wyszukiwarek internetowych zajmują się więc obecnie specjalizowane programy zwane robotami lub „pełzaczami" (web crawler). Wbrew swej angielskiej nazwie, roboty nie wędrują po sieci, lecz są po prostu czymś w rodzaju automatycznych przeglądarek internetowych (browser). Klasyczny robot to program zainstalowany na komputerze podłączonym do Internetu pobierający kolejne strony WWW. Z jednej strony na drugą robot przechodzi korzystając z odnośników – linków (links) - zamieszczonych na stronach w ramach globalnej pajęczyny WWW.

Z punktu widzenia robota przestrzeń WWW traktować można jako graf. Wierzchołkami grafu są strony HTML, a krawędziami połączenia między nimi. Zwrócić uwagę należy, że połączenia te są jednokierunkowe – gdy strona A wskazuje na stronę B, nie znaczy to, że strona B wskazuje na stronę A, choć oczywiście jest to możliwe.

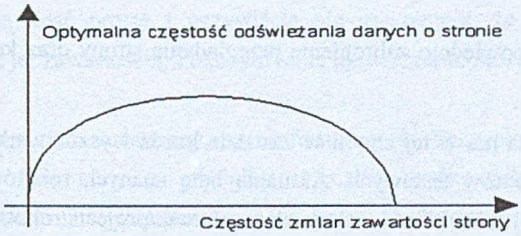


Rys. 2. Sieć WWW jako graf
Fig. 2. WWW as a graph

Najprostszym algorytmem pracy robota jest rozpocząć na jakiejś stronie z dużą ilością linków i rekurencyjnie przeglądać po kolei wszystkie znajdywane strony. Oczywiście, korzystanie z jednej tylko strony startowej prawdopodobnie nie pozwoli pokryć całej przestrzeni, należy wobec tego wybrać możliwie wiele, niezależnych od siebie, stron startowych. Jasne jest, że im więcej stron startowych zostanie wybranych, tym większą część przestrzeni pokryje robot. Jednocześnie jednak zwiększy się ilość powtórzeń – z dwóch różnych stron robot trafić może do tej samej – a co za tym idzie, spadnie wydajność. Twórcy robotów zabezpieczają się przed tym problemem przez zapamiętywanie adresów odwiedzanych stron.

Jednak nawet przejrzenie wszystkich zasobów WWW nie rozwiązuje problemu. W styczniu 2000 roku szacowano liczbę stron internetowych na 10^9 [1]. Od tej pory liczba ta z pewnością znacznie się zwiększyła. Nawet przy bardzo szybkich i efektywnych robotach (Scooter AltaVisty deklaruje przeglądanie 10 milionów stron dziennie) przejrzenie i zaindeksowanie całej sieci musi potrwać kilka miesięcy. Tymczasem średnią długość życia strony internetowej szacuje się na około cztery miesiące [2]. Jak widać, w momencie zakończenia zbierania informacji, duża część stron będzie już nieaktualna. Zawartość bazy musi więc być w trybie ciągłym odświeżana – nieaktualne linki usuwane, zmiany zawartości stron rejestrowane.

Ponieważ różnice w częstości zmian poszczególnych stron są bardzo znaczące, warto zoptymalizować działanie robota, biorąc pod uwagę przewidywaną częstość zmian każdej ze stron. Informacje na temat stron o stabilnej zawartości będą poddawane aktualizacji rzadziej niż strony, o których wiemy, że często zmieniają zawartość. Barierą w tym przypadku jest możliwa szybkość przeglądania stron zależna od przepustowości sieci. W [3] podano wykres częstotliwości zmian strony w stosunku do optymalnej częstości wizyt robota w celu odświeżenia informacji o stronie. Optymalna częstość to taka, która pozwala na zachowanie maksymalnej aktualności całego zbioru przechowywanych stron.



Rys. 3. Optymalna częstotliwość przeglądania strony w stosunku do częstości zmian jej zawartości

Fig. 3. Optimal crawl frequency to page refresh frequency

Jak widać z wykresu, przy stronach zmienianych bardzo często opłacalne staje się zmniejszenie częstości wizyt. Na wykresie nie przedstawiano konkretnych wartości liczbowych, które są różne w konkretnych przypadkach, lecz pewną stałą tendencję. Wynik jest dość zaskakujący, ale ma to swoje uzasadnienie, gdy strona ulega zmianom znacznie szybciej niż maksymalna możliwa częstość odwiedzin. Na przykład robot odwiedza stronę maksymalnie raz dziennie, a ulega ona aktualizacji co godzinę. W takim przypadku średni czas aktualności informacji o stronie w wyszukiwarce wynosi pół godziny. Oczywiste jest, że bardziej opłaca się aktualizować w tym czasie strony o większym czasie zmian, ponieważ bardziej wpływa to na zwiększenie aktualności całej kolekcji przechowywanych stron.

Ograniczenia w szybkości przeglądania stron niekoniecznie spowodowane są niedostatkami sprzętowymi twórcy robota. Wiele robotów same narzuca sobie ograniczenia w szybkości pobierania stron. Należy bowiem pamiętać, że niekontrolowane użycie robotów wiąże się z dużym niebezpieczeństwem zatkania sieci. Wystarczy, że ktoś uruchomi robota przeglądającego z dużą szybkością strony znajdujące się na pewnym serwerze, żeby serwer ten przestał odpowiadać na zgłoszenia innych użytkowników. Jest to najprostszy i jeden z najczęściej spotykanych ataków na serwery internetowe.

Tak więc od twórców robotów wymaga się przestrzegania swoistego savior-vivre'u. Należy do nich przestrzeganie postanowień **The Standard for Robots Exclusion (SRE)** [4]. Dokument ten określa, w jaki sposób serwer może ochronić swe zasoby przed przeglądaniem ich przez roboty, a co za tym idzie, przeciążeniem łącza. Należy do nich umieszczenie pliku `robots.txt`, będącego dla robota wskazówką, jak zachowywać się na danym serwerze. W pliku tym można zabronić dostępu do serwera zarówno wszystkim, jak i pewnym określonym robotom przeglądającym sieć. Inną metodą jest umieszczenie bezpośrednio w treści strony HTML zapisów w formacie META TAG:

```
<META NAME="ROBOTS" CONTENT="NOINDEX">
<META NAME="ROBOTS" CONTENT="NOFOLLOW">
```

które powodują odpowiednio zabronienie przeglądania strony oraz korzystania z linków na tej stronie.

Własnego robota ma w tej chwili w zasadzie każda wyszukiwarka internetowa. Istnieje także mnóstwo robotów testowych. Aktualną listę znanych robotów można znaleźć pod adresem internetowym http://info.webcrawler.com/mak/projects/robots/robots.html.

3.2. Przechowywanie danych (indexer)

Kolejnym problemem jest sposób przechowania zebranych przez robota danych. Robot ściąga w całości stronę i przesyła ją do programu zapisującego dane w bazie (indexer). Program taki powinien zapisać w bazie informacje na temat strony w sposób, który umożliwi jej późniejsze znalezienie. Podstawowe zagadnienie dla takiego programu, to jaką informację przechowywać.

Rozwiązanie najprostsze, ale i najbardziej obciążające bazę, to przechowywanie wszystkich znalezionych wyrazów z odnośnikami na strony, na których je znaleziono. Takie rozwiązanie stosuje na przykład popularna wyszukiwarka AltaVista [5]. Rozwiązanie to powoduje jednak konieczność przechowywania ogromnej ilości informacji, z której większość jest zupełnie nieużyteczna. Oczywiście, problem nadmiaru danych może być rozwiązany przez zwiększenie ilości pamięci masowych, a prędkość wyszukiwania przez zwiększenie mocy obliczeniowej komputerów. Jednak bardzo trudna jest w takim przypadku ocena, czego rzeczywiście dotyczy strona.

Pierwszym krokiem do uproszczenia bazy może być stworzenie predefiniowanej listy słów, które są nieużyteczne w wyszukiwaniu i nie wnoszą nic do informacji o stronie. Nikt nie będzie przecież wyszukiwał zaimków czy spójników. Pozwala to zwiększyć wydajność bazy, ale nie zwiększa w żaden sposób jakości odpowiedzi.

Innym rozwiązaniem może być przechowywanie tylko słów zawartych w specjalnych opisach strony – tak zwanych tagach (meta-tags). Jednak większość autorów stron nie respektuje meta-tagów, a wielu może wręcz specjalnie wprowadzać mylne zapisy (na przykład wpis 'windows, sterowniki' może znacząco zwiększyć ilość odwiedzeń strony, niezależnie od tego, co się na niej w rzeczywistości znajduje).

Dlatego znacznie pewniejszym rozwiązaniem jest korzystanie z rzeczywistej zawartości strony. Można indeksować na przykład tylko słowa znajdujące się w nagłówkach lub tylko pierwszych 10 linii tekstu.

Rozwiązania te są dość proste i oczywiście nie gwarantują, że opis strony będzie rzeczywiście zgodny z jej zawartością. Poza tym takie uproszczenie może spowodować utratę pewnych informacji.

Należałoby się przy tym zastanowić, czy informacja, że dane słowo znajduje się na stronie, jest rzeczywiście tym, co chciałby wiedzieć użytkownik. Zwykle przecież użytkownicy szukają stron „na jakiś temat", a nie stron „zawierających jakieś słowo". Tak więc bardziej, niż szczegółowy indeks słów, przydatna byłaby im informacja „o czym jest strona". Baza zawierająca takie informacje umożliwiałaby wyszukiwanie według tematyki

(concept-based searching) w odróżnieniu od przedstawionego wcześniej wyszukiwania według słów kluczowych (keyword searching).

Zagadnienie oceny zawartości strony nie jest sprawą banalną. Zaawansowane algorytmy bazują na teoriach lingwistycznych i sztucznej inteligencji. W interesujący sposób uprościli to zagadnienie projektanci wyszukiwarki Excite [6]. Otóż dla każdego tematu określili pewną liczbę słów kluczowych. Przeglądając stronę program indeksujący liczy częstotliwość pojawiania się słów kluczowych i sprawdza ich odległości w tekście od innych słów kluczowych z tego samego tematu. Jeśli liczba słów i odległości między nimi spełniają pewien określony poziom – program stwierdza, że strona jest właśnie na ten temat.

Wyszukiwanie tematyczne wymaga przy takim podejściu wcześniejszego utworzenia drzewa tematów i stworzenia zasad przyporządkowania do nich stron w postaci listy słów kluczowych i wymaganych odległości między nimi. Bardziej zaawansowane programy indeksujące same tworzą kategorie i rozdzielają je według właściwości. Jednak ze względu na dużą złożoność zagadnienia jest ono ciągle w trakcie rozwoju.

Patrząc na sprawę czysto matematycznie, zawartość strony można potraktować jako zestaw wartości w wielowymiarowej przestrzeni. Można sobie wyobrazić system, który na podstawie pewnej funkcji umieszcza stronę w wyliczonym miejscu tej wielowymiarowej przestrzeni i na podstawie jej analizy określa strony o zbliżonej tematyce. Jednak funkcja ta, aby działać efektywnie, musiałaby być bardzo złożona.

Sposób zapisu informacji jest kluczowy dla możliwości jej późniejszego uzyskania. Dlatego dalsze rozważania na ten temat prowadzone będą w następnym rozdziale.

3.3. Wyszukiwanie danych (searcher)

Gdy zostanie już stworzona baza indeksów do stron WWW, poważnym problemem staje się umożliwienie użytkownikom efektywnego korzystania z przechowywanych informacji. Jest to, bez wątpienia, największy problem do rozwiązania w wyszukiwarce internetowej. Problem polega na tym, aby po podaniu przez użytkownika w sposób możliwie jak najprostszy kryterium poszukiwania, wyszukiwarka dała odpowiedź jak najbardziej go satysfakcjonującą. Nie jest przy tym lepsza odpowiedź zawierająca 10 tys. linków, lecz taka, która na początku zawiera 10 odpowiedzi najważniejszych dla danego zagadnienia.

Wyszukując, użytkownik wprowadza kryterium wyboru stron. Kryterium ma najczęściej postać słowa lub listy słów, które muszą znaleźć się na stronach wynikowych. Użytkownik może także poinformować przeglądarkę o liście słów, które nie mogą znaleźć się na stronie.

W wielu przypadkach trudno jest określić, czego się poszukuje tylko za pomocą listy słów. W [7] przedstawiono historię poszukiwania informacji przez pewnego miłośnika zwierząt na temat szybkości jaguara.

Wpisanie '+jaguar +speed' dało w wyniku zupełnie nieuporządkowany zbiór stron poświęconych samochodom, grze wideo, drużynie futbolowej i serwerowi sieciowemu. Pierwsza strona na temat zwierzęcia jaguar znalazła się na 183 miejscu. Następnie spróbowano '+jaguar +speed –car' i na początku pojawiły się strony dotyczące samochodu marki Jaguar nie zawierające słowa 'car'. Kryterium '+jaguar +speed +cat' także nie dało spodziewanego rezultatu, ponieważ pojawiło się mnóstwo stron ze słowem 'cat', nie zawierających szukanej informacji.

Klasyczna wyszukiwarka na pytanie o słowo 'serce' zwróci linki do stron zawierających zarówno informacje medyczne, jak i na przykład strony walentynkowe. W odpowiedzi za to w ogóle nie pojawią się strony zawierające słowo 'kardiologia'.

W takiej sytuacji pomocne jest ograniczenie stron ze względu na ich tematykę. Aby rozwiązać ten problem, wiele wyszukiwarek pozwala na zawężenie obszaru wyszukiwania do pewnej kategorii stron. Problemy z tworzeniem tego typu drzew tematycznych zarysowano w poprzednim rozdziale. Z powodu tych problemów drzewa nie mogą być zbyt szczegółowe.

Tak więc nawet takie zawężenie przestrzeni przeszukiwanych stron powoduje wygenerowanie listy linków do wielu stron, które co prawda, dotyczą interesującego użytkownika tematu, ale w większości nie zawierają informacji, których on potrzebuje.

Podstawowym problemem staje się więc uszeregowanie stron wynikowych ze względu na ich przydatność użytkownikowi (relevancy). Najczęściej stosuje się tu pewne proste i w wielu przypadkach sprawdzające się rozwiązania. Na przykład strona jest tym wartościowsza, im wcześniej znalazło się na niej szukane słowo lub im więcej razy to słowo wystąpiło. Oczywiście, przytoczyć można wiele przypadków, w których takie uproszczenia zupełnie nie zdają egzaminu.

Wartość strony można także ocenić przez obliczenie ilości innych stron na nią wskazujących. Taką metodę zastosowali na przykład twórcy wyszukiwarki Google [9]. Obliczają oni tak wartość strony (PageRank) na podstawie wzoru:

$$PR(A) = (1-d) + d*[PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)] \quad (1)$$

gdzie:

PR(X) – wartość (PageRank) strony X,

C(X) – ilość linków wychodzących ze strony X,

T1 . . . Tn – strony wskazujące na stronę A.

d - współczynnik tłumienia.

Przy takim podejściu najwyżej klasyfikowane będą strony, na które wskazuje wiele innych, także wysoko zaklasyfikowanych stron. Wydaje się to logiczne i jak zapewniają autorzy [10], znacznie polepsza jakość odpowiedzi wyszukiwarki.

Jednak zastosowanie takiego wartościowania powoduje, że popularne strony, jak www.microsoft.com czy www.netscape.com, znajdą się na pierwszym miejscu w zbiorze wyników każdego wyszukiwania, którego warunki spełniają.

Wyszukiwanie według słów kluczowych może dawać często wyniki, wśród których nie znajdują się strony najbardziej 'reprezentatywne' dla zagadnienia, ponieważ strony te nie zawierają explicite zadanego zwrotu. Na przykład jako wynik zapytania 'car manufacturer' nie pojawią się, wydawałoby się najbardziej reprezentatywne, strony: http://www.honda.com czy http://www.ford.com, ponieważ nie zawierają one szukanego zwrotu. W [8] zaprezentowano interesujący sposób znajdowania „najlepszych" stron na dany temat z wykorzystaniem istniejących wyszukiwarek. Autorzy odpytują wyszukiwarkę tekstową podając jej zadane kryteria i zapamiętują jej pierwszych N propozycji jako zbiór stron. Następnie rozbudowują ten zbiór o strony, na które wskazują znalezione dokumenty oraz o pierwsze 50 stron wskazujących na każdy z dokumentów. Strony te są później wartościowane na podstawie ilości linków między nimi i wielu innych parametrów. Rozwiązanie to pozwala na znalezienie dokumentów o szukanej tematyce, mimo że nie znajdują się na nich bezpośrednio wpisane jako kryteria słowa.

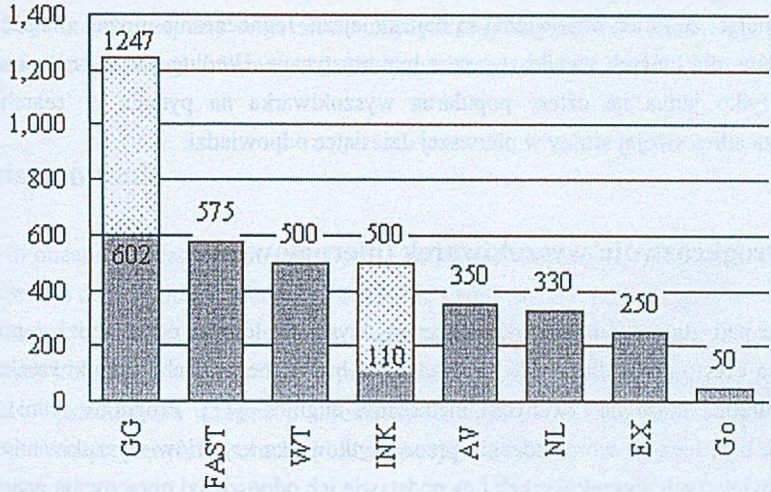
Ponieważ problem oceniania zawartości strony i jej wartości w wyszukiwaniu jest problemem złożonym, ciekawym pomysłem jest przerzucenie części pracy na użytkowników. Wyszukiwarka może rejestrować, które strony wybierali poprzedni użytkownicy poszukujący tego tematu i wartościować je ze względu na tego rodzaju „popularność". Zastosowanie takiej metody deklaruje na przykład wyszukiwarka DirectHit [13].

Poważnym i mającym coraz większe znaczenie problemem przy przeszukiwaniu zasobów WWW staje się coraz większa różnorodność typów dokumentów w sieci. Klasyczna wyszukiwarka nie jest w stanie analizować treści dokumentów zamieszczonych na przykład w formacie PDF czy PostScript. Nie da się także bezpośrednio wyszukiwać w plikach graficznych czy dźwiękowych. Faktem jest, że tego typu dokumenty są także zwykle opisane w sposób tekstowy w języku HTML – jednak sam krótki opis może nie być wystarczający przy pewnych bardziej zaawansowanych zadaniach wyszukiwania.

Coraz więcej stron internetowych generowanych jest przez serwery automatycznie w zależności od wymagań użytkowników. W ten sposób wiele informacji, dostępnych na serwerach po podaniu odpowiednich parametrów, będzie niedostępnych dla prostych robotów sieciowych – a co za tym idzie, odnośniki do nich nie znajdą się w bazach wyszukiwarek.

4. Porównywanie wyszukiwarek internetowych

Artykuł nie daje prostej odpowiedzi na pytanie „która wyszukiwarka internetowa jest najlepsza". Odpowiedź na to pytanie nie może być, oczywiście, jednoznaczna. Można tworzyć różne kryteria porównania wyszukiwarek. Najprostszym i najczęściej publikowanym kryterium jest liczba indeksowanych stron w bazach tworzonych przez wyszukiwarki. W pracy starano się jednak wykazać, że stworzenie nawet bardzo dużej bazy w żaden sposób nie rozwiązuje problemu wyszukiwania informacji. W wielu przypadkach jest dokładnie odwrotnie – im większa baza stron, tym trudniej znaleźć w niej pożądaną informację.



Rys. 4. Liczba indeksowanych stron największych wyszukiwarek
Fig. 4. Amount of pages indexed by the biggest search engines

Informacje przedstawione na rys. 4 pochodzą z listopada 2000 roku [14]. Wyjaśnienie użytych skrótów:

GG=Google, FAST=FAST, WT=WebTop.com, INK=Inktomi, AV=AltaVista, NL=Northern Light, EX=Excite, Go=Go (Infoseek).

Dwa wyniki przy wyszukiwarce Google spowodowane są faktem, że dodatkowo indeksuje ona także strony na podstawie samych linków do nich, niekoniecznie analizując ich treść. Rozszerzony wynik dla Inktomi spowodowany jest faktem, że wyszukiwarka rozdziela strony na najbardziej poszukiwane (110 milionów) oraz pozostałe, indeksując je oddzielnie.

Inne często cytowane kryterium to szybkość udzielania odpowiedzi. Ponieważ jednak szybkości wyszukiwarek są już bardzo duże, zagadnienie to staje się czysto akademickie. Dla przeciętnego użytkownika nie ma przecież żadnego znaczenia, czy czekał na odpowiedź 0,71 czy 0,54 sekundy.

Prawdziwym kryterium porównania powinna więc być łatwość znalezienia informacji. Aby to ocenić, należałoby przygotować specjalny zestaw reprezentatywnych danych do wyszukania. Nie jest to proste, zważywszy, że przeciętny użytkownik może zapytać wyszukiwarkę dosłownie o wszystko. Sprawę utrudnia dodatkowo fakt, że zawartość sieci ciągle się zmienia, a co za tym idzie, zmieniają się także zawartości baz wyszukiwarek i ich odpowiedzi na zapytania.

Wiele wyszukiwarek prezentuje przykładowe wyniki dla wybranych dziedzin, udowadniając, że to ich odpowiedzi są najtrafniejsze. Jednocześnie można znaleźć mnóstwo przykładów, dla których wyniki są wręcz humorystyczne. Według badań przeprowadzonych w [10], tylko jedna na cztery popularna wyszukiwarka na pytanie o 'search engines' umieszcza adres swojej strony w pierwszej dziesiątce odpowiedzi.

5. Drogi rozwoju wyszukiwarek internetowych

Prace nad algorytmami umożliwiającymi uzyskanie lepszej odpowiedzi z posiadanych danych są często prowadzone z wykorzystaniem baz komercyjnych wyszukiwarek. Powstała wręcz odrębna klasa tak zwanych 'metasearch engines' [12]. Programy te nie posiadają własnych baz, lecz po wprowadzeniu przez użytkownika kryteriów wyszukiwania, odpytują kilka największych wyszukiwarkek i na podstawie ich odpowiedzi opracowują wyniki.

Umożliwia to uruchomienie wyszukiwarki internetowej na praktycznie dowolnym serwerze internetowym, nie posiadającym zbyt wielu zasobów. Nie jest to, oczywiście, na rękę komercyjnym wyszukiwarkom, które zostają w ten sposób sprowadzone do narzędzia niewidocznego dla użytkownika – a co za tym idzie, mogą tracić zyski z reklam.

Komercyjne wyszukiwarki stanowią w chwili obecnej największą i najważniejszą grupę rozwiązań. Wiążą się z tym określone niebezpieczeństwa. Wyszukiwarki utrzymują się z reklam, a więc z opłacanego przez producenta informowania klientów na temat jego wyrobów czy usług. Jednocześnie jednak wyszukiwarki mogą być źródłem informacji o wyrobach firm konkurencyjnych. Jeśli jakaś firma samochodowa płaci za umieszczenie swojej reklamy na górze strony wyszukiwarki, a na zapytanie użytkownika o „tani samochód" strona tej firmy pokaże się dopiero w czwartej dziesiątce, to występuje tu wyraźny konflikt interesów użytkownika wyszukiwarki i firmy, która finansuje wyszukiwanie. Istnieje wobec tego zagrożenie, że właściciele wyszukiwarek mogą manipulować wskaźnikami jakości stron

tak, aby zadowolić swoich sponsorów. Oczywiście, jawne sprzedawanie miejsca w wynikach wyszukiwania – a były już takie próby – spotka się po prostu z protestem użytkowników, którzy przestaną używać tej strony. Jednak fakt, że większość wyszukiwarek nie udostępnia szczegółowych algorytmów oceniania jakości stron, daje im ogromne pole do niejawnej manipulacji.

Wyszukiwarki internetowe są przez niektórych uważane za zaprzeczenie idei Internetu – otwartej sieci niezależnych serwerów wymieniających się informacjami. Według nich centralizacja wszystkich zasobów w jednym miejscu sieci jest bezsensowna i w niedługim czasie przestanie być możliwa z powodu lawinowego wzrostu ilości informacji znajdującej się w Internecie. Rozwiązaniem alternatywnym mają być „wędrujące automaty" – roboty szukające w sieci informacji na zadany temat. Jak na razie rozwiązanie to jest nierealne z powodu zbyt małej przepustowości sieci. Gdyby każdy użytkownik Internetu uruchomił robota poszukującego jakiejś informacji, sieć natychmiast przestałaby działać.

6. Podsumowanie

W chwili obecnej bardzo rozwinięte są prace dotyczące opracowania sposobów zbierania i magazynowania dużych ilości informacji. Szybkość przeszukiwania potężnych baz także nie budzi zastrzeżeń użytkowników. Jest to związane głównie z coraz większą wydajnością współczesnych komputerów. Jednak dużo pracy musi zostać jeszcze wykonanej w kierunku uproszczenia wyszukiwania informacji. Problemy te wiązane są z bardzo rozbudowanymi algorytmami heurystycznymi, sztuczną inteligencją, sieciami neuronowymi i analizą leksykalną zapytań.

Jest to z całą pewnością zagadnienie kluczowe dla przyszłego rozwoju sieci informatycznych – ponieważ jak do tej pory przewaga treści nad możliwością jej wykorzystania stale się zwiększa.

LITERATURA

1. Inktomi, Inc.: Web surpasses one billion documents, http://inktomi.com, 2000.
2. Cho J., Garcia-Molina H.: The Evolution of the Web and Implications for an Incremental Crawler. Proceedings of the 26th VLDB Conference, Cairo, Egypt, 2000.
3. Cho J., Garcia-Molina H., Page L.: Efficient crawling through URL ordering. Proceedings of the 7th World-Wide Web Conference, 1998.

4. Koster M.: A Standard for Robot Exclusion. http://info.webcrawler.com/mak/projects/robots/norobots.html, 1994.
5. Digital Equipment Corporation. AltaVista Home Page. http://www.altavista.com.
6. Excite Home Page, http://www.excite.com.
7. Chakrabarti S., Dom B., Agrawal R., Raghavan P.: Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. VLDB Journal, 1998.
8. Kleinberg J.: Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999.
9. Google Home Page, http://google.stanford.edu/
10. Brin S., Page L.: The Anathomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the 7th World-Wide Web Conference (WWW7), 1998.
11. Lawrence S., Lee Giles C.: Context and Page Analysis for Improved Web Search. IEEE Internet Computing, Volume 2, Number 4, pp. 38-46, 1998.
12. Selberg E., Etzioni O. The MetaCrawler Architecture for Resource Aggregation on the Web. University of Washington, 1996.
13. DirectHit Home Page, http://www.directhit.com.
14. Search Engine Watch, http://www.searchenginewatch.com.

Recenzent: Dr inż. Arkadiusz Sochan

Wpłynęło do Redakcji 31 marca 2001 r.

Abstract

This work gathers knowledge about search engines architecture. The common problems and possible solutions are briefly described. Firstly there are some information about specific problems concerning text databases handled by search engines. Differences between these databases and classic databases are presented. Next section tells about three main parts of a complete search engine. These are crawler, indexer and searcher (Fig. 1.). Algorithms used and problems to solve in future are explained. Fig. 2 explains how web crawler treats WWW space. Fig. 3 compares frequency of refreshing data in search engine database to frequency of web pages changes. The next section shows difficulties in comparing search engines and some examples of comparison criteria (Fig. 4.). The following discusses what future work is

to be done to improve searching process. Last section is a summary, which emphasizes the importance of proper analyzing huge amounts of data from search engines' databases.