Wyszukiwanie i Przetwarzanie Informacji WWW

Wyszukiwarki WWW - Wprowadzenie

Marcin Sydow

PJWSTK

Plan wykładu

- Wprowadzenie
- Rola i funkcjonalność wyszukiwarek
- Czym wyszukiwanie w WWW różni się od wyszukiwania w korpusach tekstowych
- Moduły typowej wyszukiwarki
- Wyzwania techniczne
- Inne modele wyszukiwarek
- Podsumowanie

Web Dzisiaj

Rozmiar WWW:
dziesiątki miliardów stron (wg. worldWideWebSize.com na 30.09.2009)

kilkanaście miliardów indeksowalnych dokumentów

Ilość użytkowników WWW:
około 300.000.000 (wg. Nielsen/NetRatings 2007)
około 700.000.000 unikalnych użytkowników (comScore World Metrix, 2006.03)

kilkaset milionów użytkowników

Najpopularniejsze adresy URL

Spośród kilkunastu miliardów - jakich jest 5 najpopularniejszych witryn na świecie?

Najpopularniejsze adresy URL

Spośród kilkunastu miliardów - jakich jest 5 najpopularniejszych witryn na świecie?

- Google.com
- Facebook.com
- YouTube.com
- Yahoo.com
- Live.com

(wg. alexa.com 3.03.2010, kolejność bywa różna wg. różnych kryteriów)

3 z pięciu to wyszukiwarki, tzw. "Wielka Trójka", a 2 pozostałe należą do wyszukiwarek. Dlaczego wyszukiwarki są najpopularniejszymi serwisami?

Wyszukiwarki - motywacja

- WWW jest **największym źródłem danych i informacji**
- Informacji jest za dużo dla pojedynczego człowieka
- Cały ten ocean informacji byłby **bezużyteczny** bez narzędzia umożliwiającego sensowny dostęp
- **Dlatego:** Wyszukiwarki stanowią dzisiaj punkt wyjścia użytkowników WWW

Fakty: 256.000.000 ludzi skorzystało z wyszukiwarki w grudniu 2006 (wg. Nielsen/NetRatings)

Wyszukiwarkowe Zoo - nie tylko Google!

Obecnie istnieje kilkaset działających wyszukiwarek, nie licząc specjalnych, działających w przeszłości (przejętych, etc.). Oto niektóre z nich:

Wyszukiwarkowe Zoo - nie tylko Google!

Obecnie istnieje kilkaset działających wyszukiwarek, nie licząc specjalnych, działających w przeszłości (przejętych, etc.). Oto niektóre z nich:

(niektóre) globalne (alfabetycznie): Ask.com (dawniej Ask Jeeves); Bing (dawniej MSN Search i Live Search); Cuil; Duck Duck Go; Gigablast; Google; Kosmix; WolframAlpha; Vivisimo; Yahoo! Search; Yebol, etc...

Polska: Netsprint.pl (mniej popularne: Szukacz, Szook, Gooru; nieaktywne: Emulti, NEToskop, Sieciowid, etc...)

Wyszukiwarkowe Zoo - nie tylko Google!

Obecnie istnieje kilkaset działających wyszukiwarek, nie licząc specjalnych, działających w przeszłości (przejętych, etc.). Oto niektóre z nich:

(niektóre) globalne (alfabetycznie): Ask.com (dawniej Ask Jeeves); Bing (dawniej MSN Search i Live Search); Cuil; Duck Duck Go; Gigablast; Google; Kosmix; WolframAlpha; Vivisimo; Yahoo! Search; Yebol, etc...

Polska: Netsprint.pl (mniej popularne: Szukacz, Szook, Gooru; nieaktywne: Emulti, NEToskop, Sieciowid, etc...)

(niektóre) lokalne: Accoona, China/US; Alleba, Philippines; Ansearch, Australia/US/UK/NZ; Baidu, Sogou, Sohu: China; Daum, Korea; Goo, Japan; Guruji.com, India; Leit.is, Iceland; Maktoob, Arab World; Onkosh, Arab World; Miner.hu, Hungary; Najdi.si, Slovenia; Naver, Korea; Rambler, Russia; Rediff, India; SAPO, Portugal/Angola/Cabo Verde/Mozambique; Search.ch, Switzerland; Sesam, Norway, Sweden; Seznam, Czech Republic; Walla!, Israel; Yandex, Russia; ZipLocal, Canada/US;

Wyszukiwarkowe Zoo - nie tylko Google!

Obecnie istnieje kilkaset działających wyszukiwarek, nie licząc specjalnych, działających w przeszłości (przejętych, etc.). Oto niektóre z nich:

(niektóre) globalne (alfabetycznie): Ask.com (dawniej Ask Jeeves); Bing (dawniej MSN Search i Live Search); Cuil; Duck Duck Go; Gigablast; Google; Kosmix; WolframAlpha; Vivisimo; Yahoo! Search; Yebol, etc...

Polska: Netsprint.pl (mniej popularne: Szukacz, Szook, Gooru; nieaktywne: Emulti, NEToskop, Sieciowid, etc...)

(niektóre) lokalne: Accoona, China/US; Alleba, Philippines; Ansearch, Australia/US/UK/NZ; Baidu, Sogou, Sohu: China; Daum, Korea; Goo, Japan; Guruji.com, India; Leit.is, Iceland; Maktoob, Arab World; Onkosh, Arab World; Miner.hu, Hungary; Najdi.si, Slovenia; Naver, Korea; Rambler, Russia; Rediff, India; SAPO, Portugal/Angola/Cabo Verde/Mozambique; Search.ch, Switzerland; Sesam, Norway, Sweden; Seznam, Czech Republic; Walla!, Israel; Yandex, Russia; ZipLocal, Canada/US;

Oprócz tego: **meta-wyszukiwarki** (np. Dogpile), **wyszukiwarki open-source** (np. Egothor), **wyszukiwarki specjalistyczne** (np. Lexis), **wyszukiwarki portalowe** (np. Amazon), etc.

Historia wyszukiwania w sieci w pigułce...

1973 DARPA, 1980 FTP (anonimowe konta FTP, brak jakiegokolwiek wyszukiwania trzeba było znać dokładny adres i nazwę pliku(!), WWW 1989 w CERN (European Organisation for Nuclear Research, zał. 1954 koło Genewy) - Tim Berners-Lee, początkowo tylko do komunikacji naukowców, w 1991 otwarty na świat, Archie 1989 (przeszukiwanie FTP), Gopher 1991 (j.w.), www wanderer 1993 (pomiar WWW), Aliweb, jumpStation, WWW Worm 1994 (pierwszy system wyposażony w indeks), webCrawler (pierwszy pełny indeks tekstowy), 1995 Lycos (CMU, 60M stron, komercjalizacja w 1996), Infoseek (1994), Hotbot (1996), 1997 Ask Jeeves, Northern Light, OpenText - płatne "rankingi"[1]

[1] Zagadka: a jak jest np. w Amazon?

historia, cd...

Alta Vista (DEC, duże zasoby obliczeniowe - “Alpha servers”, po kilku zmianach ostatecznie zakupiona w 2003 przez Yahoo!), 1994 Yahoo! (David Filo, Jerry Yang, “Yet another hierarchical officius oracle”), 1998 Google (nazwa od “Googol”: ’1’ i sto zer), Yahoo: 2002 zakupiło Inktomi a w 2003 AltaVista, w 2004 uruchamia własny system wyszukiwawczy (do tej pory przez Google), AOL kupuje Excite (które zakupiło WebCrawler w 1997) ale od 2002 zaczyna korzystać z usług Google, 2005 Microsoft uruchamia własną wyszukiwarkę MSN Search (do tej pory przez technologię Inktomi będącą własnością Yahoo!), Ask Jeeves 2001 kupuje Teoma, a w 2005 zakupiony przez InterActiveCorp (od teraz: Ask.com)

Co powinna robić wyszukiwarka?

Zwrócić informacje zawarte w WWW zgodne z potrzebą informacyjną użytkownika

Najpopularniejszy dzisiaj wariant:

- wejście: wyrażenie potrzeby informacyjnej - (np. zapytanie "boolowskie")
- wyjście: prezentacja informacji - (np. lista linków do dokumentów zawierających dane słowa)

Ten wariant wcale nie jest doskonały - użytkownik oczekuje informacji a nie listy dokumentów.
(wyjątkiem są tzw. "zapytania nawigacyjne" (ang. navigational queries))

Możliwe są inne niezliczone warianty.

Wyszukiwarki "boolowskie"

Zadanie jest "proste":
zwrócić dokumenty WWW zawierające dane słowa kluczowe

- odruchowo używane wielokrotnie w ciągu dnia
- minimalistyczny interfejs

w istocie **bardzo skomplikowane** systemy

ilustrują pełne spektrum zagadnień algorytmicznych, analizy danych, inżynieryjnych, technologicznych, ...

Wyszukiwarki a klasyczne IR - specyfika WIR

Na pierwszy rzut oka zadanie wyszukiwarki nie różni się bardzo od klasycznego systemu wyszukiwania tekstowego:

- jest korpus dokumentów (po odrzuceniu znaczników html mógłby to być korpus tekstowy)
- jest zapytanie "boolowskie"
- należy zwrócić dokumenty zawierające słowa kluczowe

A jednak, to podobieństwo jest tylko pozorne. W istocie wyszukiwanie w WWW jest na tyle specyficzne, że klasyczne systemy IR nie nadają się do tego celu.

Po pierwsze, nie istnieje tu gotowy "korpus dokumentów". Należy go dopiero zebrać z WWW za pomocą specjalnego, skomplikowanego oprogramowania sieciowego (tzw. crawler).

Poza tym sam "tekst" w WWW sprawia rozmaite problemy.

WWW - problemy z tekstem

Klasyczne, tekstowe techniki IR sprawiają problemy w przypadku WWW:

- Problem skali (ogromny "korpus")
- Problem **braku samo-opisu**
  (np. zapytanie: "japoński producent samochodów")
- Problem różnorodności
- Problem nierównej jakości
- Zaszumienie, błędy, etc.
- Tekst - łatwy do spamowania

WWW - rozwiązanie problemów IR

WWW z jednej strony **stwarza problemy** dla klasycznego IR. Z drugiej strony, **stwarza możliwości** ich obejścia dzięki istnieniu dodatkowych źródeł informacji:

- społeczny aspekt publikowania w WWW (linki)
- tekst odnośników (ang. anchor text)

To są mocne narzędzia:

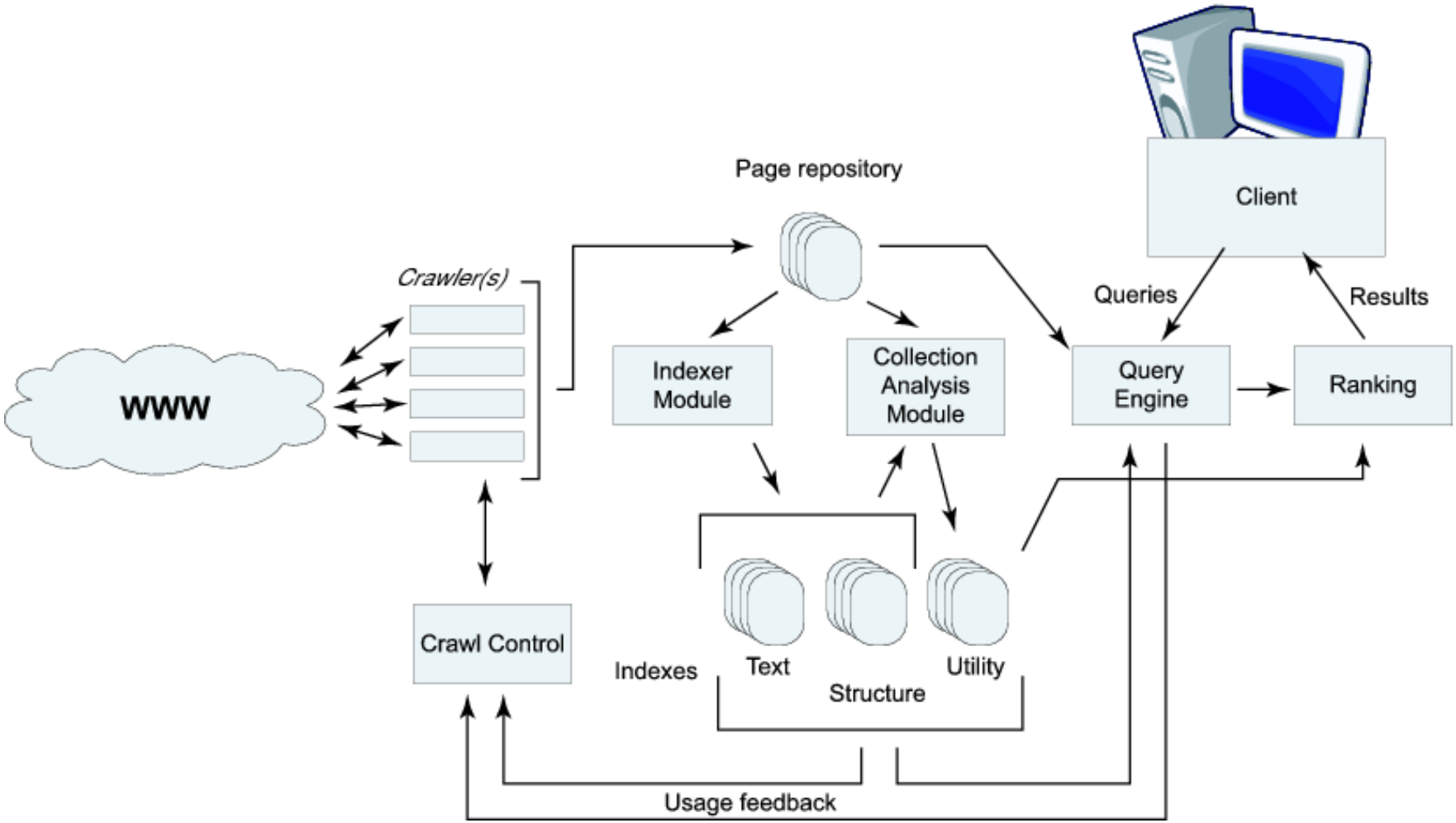
- ominięcie problemu braku samo-opisu
- dokumenty nietekstowe
- dokumenty o nieznanych formatach
- dokumenty nieściągnięte

Dodatkowo: nazwa hosta, domeny, pliku, głębokość ścieżki, ilość dokumentów na hoście, ...

Moduły wyszukiwarki

- Moduł zbierający (ang. **Crawler**)
  - podążaj po linkach i ściągaj dokumenty
- **Repozytorium**
  - składuj ściągnięte dokumenty - trwałość, dostęp
- **Indeks**
  - zapisz które słowo występuje w jakim dokumencie
- **System Rankingowy**
  - jakie informacje dobrze pasują do zapytania użytkownika?
  - jakie informacje są wartościowe same w sobie?
- **Moduł prezentacji**
  - znajdź dobrą formę wizualizacji wyników
- **Obsługa**
  - obsłuż zapytania, znajdź strony, wyświetl wyniki

Schemat Ogólny Architektury Wyszukiwarki



(wg "Searching the Web" A.Arasu, et al.)

Zbieranie dokumentów (crawler)

Idea jest prosta:

kolejka (priorytetowa) adresów URL

praca w cyklach:

- pobierz URL z kolejki
- ściągnij odpowiadający mu dokument
- sparsuj aby wyciągnąć hiper-linki
- wrzuć hiper-linki do kolejki
- zachowaj dokument w repozytorium
- powtarzaj to bez końca...

Polska: np. dla 85 mln dokumentów: 34 dokumenty/s codziennie przez cały miesiąc

kolejka jest na ogół rozproszona, a operacja współbieżna

Zbieranie dokumentów, c.d.

Wykonanie jest już mniej proste:

- Etyka:
  - robots.txt
  - interwał np. 5s dla danego hosta (zamiast 34dok/s: 170 jednocześnie aktywnych połączeń (w skali świata (x100): **15 000** aktywnych połączeń)
- połączenia z setkami tysięcy nieznanych serwerów (DNS, kodowanie, błędy html, błędy sieciowe, etc.)
- rozproszony, wielowątkowy system sieciowy, architektura odporna na błędy i przeciążenie, trudne zagadnienia z dziedziny algorytmów i struktur danych
- pułapki na crawlery, e-maile od webmasterów...

Indeks

Centralna struktura danych systemu

Dla każdego słowa: gdzie ono występuje (przy czym jest wiele "kontekstów")

W momencie zapytania: pobiera listy i odpowiednio je łączy (intensywne obliczeniowo, zaawansowane ASD)

Przygotowanie indeksu: bardzo kosztowne obliczeniowo

- wykonywane cykliczne (co najmniej miesięcznie)
- również zaawansowane ASD

Szukanie igły w stogu siana - Ranking

Przeciętne zapytanie: **tysiące zwróconych** dokumentów

Możliwości użytkownika: **kilkanaście obejrzanych** dokumentów

Szukanie igły w stogu siana - Ranking

Przeciętne zapytanie: **tysiące zwróconych** dokumentów

Możliwości użytkownika: **kilkanaście obejrzanych** dokumentów

Jak wybrać **na początek listy te kilkanaście najlepszych** spośród tysięcy?

Szukanie igły w stogu siana - Ranking

Przeciętne zapytanie: **tysiące zwróconych** dokumentów

Możliwości użytkownika: **kilkanaście obejrzanych** dokumentów

Jak wybrać **na początek listy te kilkanaście najlepszych** spośród tysięcy?

Rozwiązaniem jest: System Rankingowy

Systemy rankingowe istniały od lat w IR, ale nie były idealne w przypadku WWW
(rewolucja wyszukiwarkowa AD 1998)

Ranking

Najpilniej strzeżone tajemnice wyszukiwarek (decydują o **jakości wyników**)

Dokumentowi przyporządkowana jest wartość (ang. score) i wyniki są posortowane po tej wartości

Wiele składowych:

- analiza tekstu (zawartość, URL, meta, ...)
- analiza tekstu odnośników (ang. anchor text)
- analiza struktury linków
- analiza logów, ruchu internetowego, ...

Analiza Tekstu

Dziedzictwo po IR (jednorodne kolekcje)

Fazy:

- oczyszczanie (odkodowanie, jakie symbole, kapitalizacja)
- usuwanie niby-słów (ang. stop-words) (Polski: “ale”, “lub”, etc. Ale uwaga na tematykę)
- lematyzacja (w angielskim: np. algorytm Portera)
- wybór istotnych cech (ang. feature selection)
- obliczenie reprezentacji (multizbiór słów - ang. “bag of words”, wektor bitowy, model probabilistyczny, indeks, etc.) - zależy od zadania i modelu

Tekst a ranking

- statystyki (np. tf-idf)
- pozycja w tekście
- pozycja w kontekście (URL, meta, title, anchor, etc.)
- meta-znaczniki
- znaczniki prezentacji (rozmiar, pogrubienie nagłówek)

Obsługa zapytania użytkownika: operatory

Podobnie jak w klasycznych tekstowych systemach IR, zapytanie składa się ze słów kluczowych oddzielone operatorami algebraicznymi: AND (domyślny), OR, NOT.

Oprócz tego obecny jest "operator frazy" oraz czasami operatory bliskości (rzadko używane)

Operatory te (i inne) są na ogół dostępne w wyszukiwarkach przez interfejs "wyszukiwania zaawansowanego" (gdzie wypełnia się odpowiednie pola formularza). Można jednak wpisywać je bezpośrednio przy zastosowaniu się do odpowiedniej **składni**.

Pozostałe operatory

Współczesne wyszukiwarki na ogół udostępniają jeszcze szereg dodatkowych, specjalistycznych operatorów, pozwalających na zawężenie wyników tylko do dokumentów o określonych np.:

- formatach
- datach powstania
- zawierających poszukiwane wyrazy w rozmaitych "kontekstach"
- hostach lub domenach, na których występują
- dla których poszukiwane wyrazy występują w zadanej maksymalnej odległości

Szczegóły zależą od poszczególnych wyszukiwarek i są zwykle opisane w interfejsie "wyszukiwanie zaawansowane" danej wyszukiwarki.

Normalizacja Zapytań

Zapytanie zawierające wiele różnych operatorów logicznych jest na ogół normalizowane przez moduł obsługi zapytania przed rozpoczęciem obliczania wyników.
Na ogół usuwane są zbędne człony, znaki, niby-wyrazy, i całość sprowadzana jest do postaci koniunkcji.

Moduł Prezentacji

Wyniki, po obliczeniu należy jeszcze **zaprezentować**:

- Zwykle informacje pochodzą z wielu różnych maszyn
- Dochodzą reklamy, linki sponsorowane, które należy dopasować
- Dochodzą informacje kontekstowe (np. o osobach)

Często występują elementy dodatkowe:

- auto-korekcja zapytania
- grupowanie wyników
- sugerowanie następnego zapytania
- dodatkowe informacje (np. “popularna strona”, etc.)

Wymagania czasowe

Podsumujmy obsługę pojedynczego zapytania:

- parsowanie zapytania
- rozproszenie obliczeń
- łączenie list w indeksie
- obliczenie rankingu
- połączenie wyników zgodnie z rankingiem
- wyświetlenie wyników
- obliczenie reklam, korekcji, podpowiedzi, ...

Cały ten cykl **musi** być obsłużony w **ułamku sekundy** dla indeksu odpowiadającego **dziesiątkom TB**

Wymagania czasowe

Podsumujmy obsługę pojedynczego zapytania:

- parsowanie zapytania
- rozproszenie obliczeń
- łączenie list w indeksie
- obliczenie rankingu
- połączenie wyników zgodnie z rankingiem
- wyświetlenie wyników
- obliczenie reklam, korekcji, podpowiedzi, ...

Cały ten cykl **musi** być obsłużony w **ułamku sekundy** dla indeksu odpowiadającego **dziesiątkom TB**

ile przychodzi zapytań w ciągu sekundy?

Przykład

Załóżmy **500.000.000 zapytań dziennie** w skali globalnej (wg. Google, 2005)

Załóżmy, że największa wyszukiwrka dostaje ok 50% tego ruchu (46% G, 23%Y, 11%M (NetRatings, 2005),(UK, grudzień 2006: 77%, 8%, 5%))

dla pojedynczej wyszukiwarki oznacza to:

- 230M zapytań dziennie,
- czyli ponad **2500 zapytań na sekundę**

dane oszacujmy z grubsza na 80 TB (dla 8G dokumentów tekstowych)

Przykład

Załóżmy **500.000.000 zapytań dziennie** w skali globalnej (wg. Google, 2005)

Załóżmy, że największa wyszukiwrka dostaje ok 50% tego ruchu (46% G, 23%Y, 11%M (NetRatings, 2005),(UK, grudzień 2006: 77%, 8%, 5%))

dla pojedynczej wyszukiwarki oznacza to:

- 230M zapytań dziennie,
- czyli ponad **2500 zapytań na sekundę**

dane oszacujmy z grubsza na 80 TB (dla 8G dokumentów tekstowych)

Ile sprzętu trzeba aby to obsłużyć?

Jak wygląda wyszukiwarka od kuchni...

Zastanówmy się jak fizycznie "wygląda" wyszukiwarka.

Przy obciążeniach obliczeniowych tej skali typowa globalna wyszukiwarka potrzebuje naprawdę ogromnych zasobów obliczeniowych:

Np. dane techniczne dotyczące sprzętu utrzymywane są w ścisłej tajemnicy ale typowa maszynownia Google to **klaster rzędu dziesiątek tysięcy** pracujących bez przerwy serwerów, na których uruchomione jest specjalne oprogramowanie (włączając w to specjalnie zmodyfikowany system operacyjny ("prywatna" wersja Linuxa)), z oddzielnym zasilaniem, instalacją przeciwpożarową, który na okrągło jest rozbudowywany, wymieniane są zużyte części, etc. Przy czym takich klastrów jest wiele i są one rozproszone geograficznie (np. w Północnej Karolinie, Oregon, Kaliforni, Holandii, ...).

Innymi słowy, miejsca gdzie obliczane są odpowiedzi na nasze zapytania to po prostu gigantyczne centra obliczeniowe, co nie jest oczywiste gdy patrzymy na minimalistyczny interfejs wyszukiwarki.

Wyszukiwarki - Wyzwania

Przy takiej skali zadań wyszukiwarki stoją przed ekstremalnymi problemami:

- algorytmicznymi (jak to szybko liczyć)
- programistycznymi
- architekturalnymi
- sprzętowymi
- finansowymi
- etycznymi (co można a czego nie?)
- ...fizycznymi (np. jak chłodzić)

Rozszerzenia Podstawowej Funkcjonalności

- Personalizacja
- Autokorekta Zapytań (ang. Query Correction)
- Podpowiadanie Zapytań (ang. Query Suggestion)
- Rozpoznawanie "typu" obiektu (np. osoba, firma)
- Mapy Dokumentów
- Grupowanie Dokumentów (Vivisimo, Carrot2)
- Znajdowanie Materiałów Podobnych

Modele Wyszukiwarek

Najpopularniejszym typem są globalne wyszukiwarki ogólne, ale można wymienić też inne modele:

- Wyszukiwarki Portalowe i Intranetowe (w "pół drogi" pomiędzy klasycznym IR a WIR. Specyficzne aspekty: kontrolowane, brak spamu, specyficzna struktura, mniejsza skala)
- Wyszukiwarki Tematyczne
- Wyszukiwarki Wiadomości

Powyższe modele różnią się znacznie co do założeń i zasady działania od modelu podstawowego

Z punktu widzenia architektury systemu istotną alternatywą są też wyszukiwarki P2P, które są jednak dopiero w fazie rozwojowej.

Interesujący, odrębny model stanowią też tzw. **"meta-wyszukiwarki"** (np. `dogpile.com`)

Co wypada wiedzieć po tym wykładzie:

1. Rola i funkcjonalność wyszukiwarek
2. Czym wyszukiwanie w WWW różni się od wyszukiwania w korpusach tekstowych
3. Moduły wyszukiwarki i ich funkcje
4. Orientacyjne liczby dotyczące wyszukiwarek
5. Wyzwania wyszukiwarek
6. Inne modele wyszukiwarek

Dziękuję za uwagę