Zarządzanie treścią (CMS)

Wyszukiwanie treści

Zagadnienia

- Wyzwania Internetu
- Problemy techniczne
- Paradygmaty wyszukiwania
- Szum informacyjny
- Inteligentne wyszukiwarki internetowe
- Syntagmatyka
- Podsumowanie

Wykorzystano materiały prof. K. Subiety

Wyzwania Internetu

- Zmniejszenie popularności tradycyjnej gazety na rzecz gazet internetowych.
- Duże zmiany w telefonii.
- Duże zmiany w organizacji i metodach nauczania.
- Duże zmiany w organizacji i kontroli pracy (umożliwienie pracy w domu przy zapewnieniu pełnej kontroli ze strony pracodawcy).
- Handel, biznes, administracja oparte na Internecie.

Stan obecny zasobów WWW (1)

- Wykładniczy wzrost ilości stron WWW
- Pogorszenie się jakości prezentowanych na nich informacji (duplikaty, nieaktualność, banały - góry śmieci!)
- Nie zawsze najlepsze dostosowanie metod wyszukiwawczych do potrzeb użytkowników i ich możliwości intelektualnych.

Stan obecny zasobów WWW (2)

- „Głęboki Web” – większość informacji dostępnych przez Web nie ma formy płaskich stron HTML, lecz jest przechowywana w bazach danych i na bieżąco montowana w postaci strony HTML w odpowiedzi na zapytanie użytkownika.
- To powoduje, że zliczanie stron HTML jest całkowicie nieadekwatne – w bazach danych siedzi praktycznie nieskończona kombinacja informacji, które mogą być zaprezentowane jako strony Web.

Stan obecny zasobów WWW (3)

- Jakkolwiek większość popularnych standardów tekstowych została zaabsorbowana przez popularne wyszukiwarki, istnieją też takie formaty jak audio, grafika, wideo, które są nierozpoznawalne w sieci i muszą być zaindeksowane explicite lub kontekstowo.
  - Jest to pracochłonne.
  - Metody sztucznej inteligencji nie zawsze są wystarczająco zaawansowane.

Stan obecny zasobów WWW (4)

- Bardziej inteligentne metody wyszukiwania, bazujące na inżynierii lingwistycznej, są mało skuteczne wobec rozmiaru zasobów Web.
- Użytkownika nie interesuje informacja jako taka, lecz informacja niezbędna dla rozszerzenia jego wiedzy lub podjęcia decyzji.
  - **Relewancja**: informacja odpowiada formalnie zapytaniu użytkownika.
  - **Trafność** (pertinency): informacja odpowiada potrzebie użytkownika.

Zmiany jakościowe w organizacji Web

- Trwają pracę nad tzw. semantycznym Webem (*semantic web*), który będzie Webem na wzór dobrze zorganizowanej bazy danych.
- Jako narzędzie strukturalizacji proponuje się XML i w tym kierunku idzie ogromny strumień R&D.
- XML jest dobry jako podstawa standaryzacji różnorodnych protokołów wymiany informacji, ale jest bardzo ograniczony jako model danych.

Zmiany jakościowe w organizacji Web

- XML – c.d:
  - Mizerna podstawa semantyczna XML-owego modelu danych daje efekt piramidy stojącej na czubku, która wymaga różnorodności „podpórek” .
  - Te podpórki wprowadzają dodatkowy chaos do technologii dookoła-Webowych, powodując monstrualny (i niepotrzebny) rozrost terminologii, pojęć i dokumentacji. Mimo to, pozostają nadal istotne ograniczenia.
  - Wydaje się, że XML nie utrzyma się jako technologia rządząca środkiem systemów zarządzania treścią. Świat komercyjny ma złudzenia co do roli XML jako modelu danych. XML pozostanie tylko środkiem wymiany informacji.

Problemy techniczne Web

- Klasyfikacja/kategoryzacja zasobów Webu: potencjalny standard klasyfikacyjny na wzór klasyfikacji dziesiętnej.
  - Ilość haseł tematycznych szacuje się na 50 000, ale to może być niewystarczające.
  - Prawdopodobnie konieczne będzie powołanie międzynarodowej organizacji zajmującej się bieżącą standaryzacją haseł tematycznych (rozrost haseł).
  - System kategoryzacji musi być wspomagany przez narzędzia automatycznego indeksowania dokumentów znajdujących się w zasobach Web.
- Ze względu na ogrom Webu powyższy standard raczej nie powstanie.
- Problem wielojęzyczności Webu (krytyczny dla zastosowań B2B oraz w pewnym stopniu dla B2C).

Usługi Internetowe

- Popularnie, Internet jest w Polsce kojarzony z WWW (protokół HTTP).
  - W tej chwili obejmuje on jednak ogromną liczbę innych usług. Wszystkie są oparte na tym samym protokole TCP/IP.
    - Email, News (Usenet),
    - FTP, SFTP
    - ICQ (ułatwiający kontakt w internecie), GG
    - Telnet, SSH (Secure Shell)
    - IRC
    - VoIP (np. Skype)
    - RSS
- Nie jest wykluczone, że może w każdej chwili pojawić się zupełnie nowa usługa, która zdominuje pewien sektor obecnie opanowany przez WWW.

Oprogramowanie dla Internetu

- Serwery WWW (Web Servers) – udostępniają klientom WWW (internautom) serwisy WWW:
  - Apache, Microsoft IIS Server, ...
- Przeglądarki WWW (Web browsers) – pozwalają ściągać i wyświetlać pliki ściągnięte z zasobów znajdujących się pod kontrolą serwerów WWW, najczęściej pliki w formacie HTML, ale nie tylko.
  - Chrome, Internet Explorer, Firefox, Safari, Opera, ...
- Dedykowanych serwerów i wyszukiwarek, szczególnie dla technologii P2P.
  - BitTorrent
  - …

Schemat architektury dla Internetu

Przeglądarka WWW

Przeglądarka WWW

Sieć
Internet

Serwer WWW

Serwer WWW

Serwer WWW

Serwer WWW

Strony HTML,
inne pliki,
bazy danych

Strony HTML,
inne pliki,
bazy danych

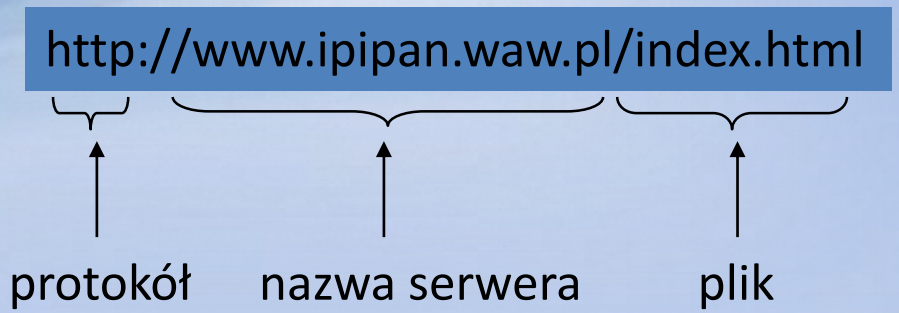
Strony HTML,
inne pliki,
bazy danych

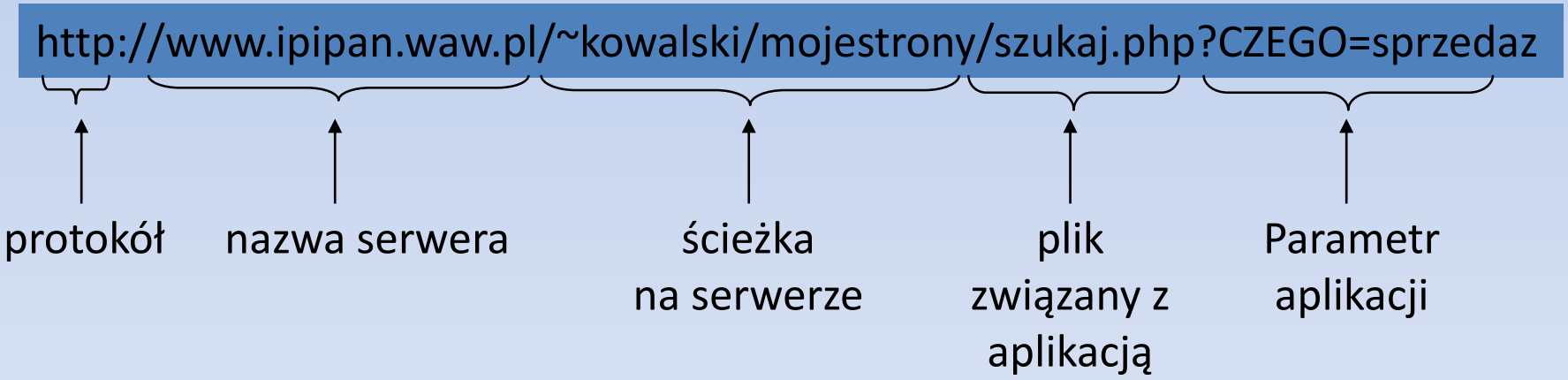
Strony HTML,
inne pliki,
bazy danych

Adresy w sieci

- Każdy zasób w sieci jest dostępny pod adresem (URL, URI), który składa się z:
  - Nazwy protokołu: http, ftp, ...
  - Adresu serwera WWW, czyli cztero-liczbowego adresu IP. Adres ten jest zwykle odwzorowany do postaci nazwowej przez specjalne serwery zwane DNS.
  - Ścieżki na systemie plików danego serwera WWW.
  - Konkretnej nazwy pliku.
  - Listy parametrów, która jest przekazywana do aplikacji związanej z w/w nazwą pliku.
- Najnowsze podejścia operują adresami „logicznymi”, które nie mają bezpośredniego przełożenia na system plików.

Przykłady URL





Nowe adresy IP

- Tradycyjne są 4-ro bajtowe i ich zapas wyczerpuje się. Są w stanie potencjalnie zaadresować 2^{32}, czyli ok. 4 miliardy serwerów, ale dodatkowe ograniczenia powodują, że jest to liczba znacznie mniejsza.
- Nowe adresy IP (IPv6) są 16-to bajtowe, co oznacza potencjalną możliwość **zaadresowania każdego centymetra kwadratowego kuli ziemskiej**, np.: 2001:0db8:85a3:0042:1000:8a2e:0370:7334

Nowe adresy IP (2)

- Nowy protokół będzie znacznie lepiej uwzględniał kwestie bezpieczeństwa.
- Ma bezkolizyjnie współpracować ze starą wersją.
- Jak na razie dość słabe wykorzystanie: ok. 2% (Roberts, Phil (24-09-2013). "IPv6 Deployment Hits 2%, Keeps Growing“.)

Wyszukiwarki stron WWW

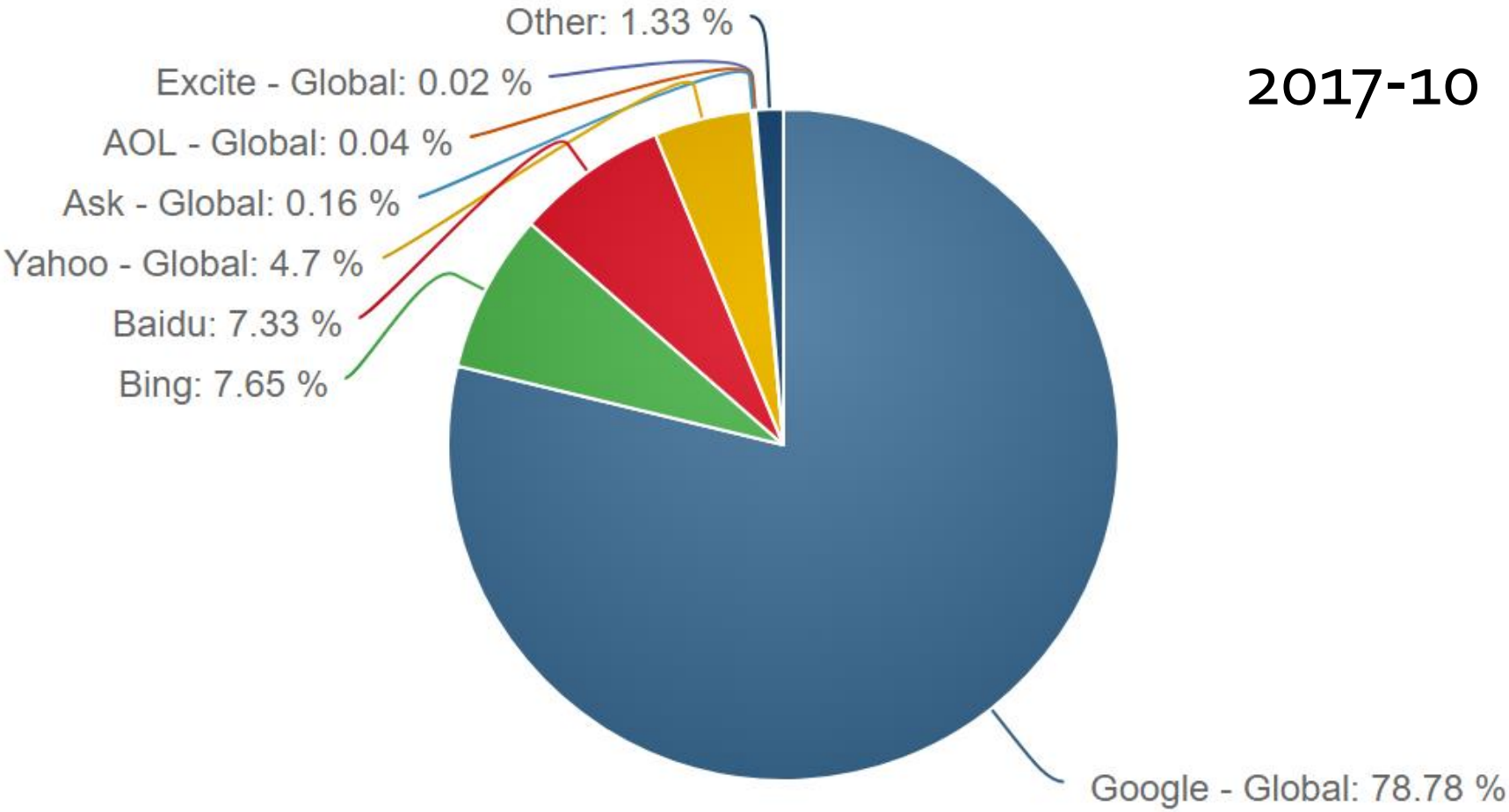
- Ogromny rozmiar zasobów Webu powoduje konieczność korzystania z wyszukiwarek.
- Na rynku pozostała niewielka liczba wyszukiwarek (Google, Bing, ...), które się sprawdziły i mają swoich wiernych klientów. Pozostałe wyszukiwarki przegrały walkę o rynek.
- Wyszukiwarki w zasadzie wyszukują zadane słowa kluczowe w pełnym tekście dokumentów znajdujących się w zasobach Web.
- Jak na razie, nie sprawdziły się nadzieje na włożenie istotnej „inteligencji” do wyszukiwarek. Są to dość proste mechanizmy.

Historia wyszukiwarek

Źródło: http://en.wikipedia.org/wiki/Search_engines

Rok	Nazwa	Zdarzenie
1993	W3Catalog	Start
	Aliweb	Start
	JumpStation	Start
1994	WebCrawler	Start
	Go.com	Start
	Lycos	Start
1995	AltaVista	Start
	Daum	Założenie firmy
	Open Text Web Index	Start
	Magellan	Start
	Excite	Start
	SAPO	Start
	Yahoo!	Start jako katalog
1996	Dogpile	Start
	Inktomi	Start
	HotBot	Założenie firmy
	Ask Jeeves	Założenie firmy
1997	Northern Light	Start
	Yandex	Start
1998	Google	Start
	MSN Search	Start
1999	AlltheWeb	Start
	GenieKnows	Założenie firmy
	Naver	Start
	Teoma	Założenie firmy
	Vivisimo	Założenie firmy
2000	Baidu	Założenie firmy
	Exalead	Założenie firmy
2002	Inktomi	Wykupienie
2003	Info.com	Start

Rok	Nazwa	Zdarzenie
2004	Yahoo! Search	Start jako wyszukiwarka
	A9.com	Zamknięcie
	Sogou	Start
2005	Ask.com	Start
	GoodSearch	Start
	SearchMe	Założenie firmy
2006	wikiseek	Założenie firmy
	Quaero	Założenie firmy
	Ask.com	Start
	Live Search	Start jako MSN Search
	ChaCha	Start
	Guruji.com	Start
2007	wikiseek	Zamknięcie
	Sproose	Zamknięcie
	Wikia Search	Start
	Blackle.com	Start
2008	Powerset	Wykupienie przez Microsoft
	Picollator	Zamknięcie
	Viewzi	Zamknięcie
	Cuil	Start
	Boogami	Start
	LeapFish	Beta Start
	Forestle	Start
	VADLO	Start
	Duck Duck Go	Start
2009	Bing	Start jako Live Search
	Yebol	Start wersji beta
	Mugurdy	Zamknięcie
	Goby	Start
2010	Yandex	Start wersji angielskiej
	Cuil	Zamknięcie
	Blekko	Start wersji beta
	Viewzi	Zamknięcie
	Yummly	Start



2017-10

http://www.netmarketshare.com/

Popularność wyszukiwarek (2)

Top 10 Search Providers for August 2009, Ranked by Searches (U.S.)			
Search Provider	Searches (000)	Month-on-Month Growth (%)	Share of Searches (%)
Total	**10,812,734**	**2.9**	**100.0**
Google	6,986,580	2.6	64.6
Yahoo	1,726,060	-4.2	16.0
MSN/WindowsLive /Bing	1,156,415	22.1	10.7
AOL	333,231	1.8	3.1
Ask.com	186,270	2.9	1.7
My Web	128,432	0.5	1.2
Comcast	50,328	-21.6	0.5
Yellow Pages	37,923	2.7	0.4
NexTag	31,830	0.4	0.3
Local.com	16,314	2.9	0.2

Źródło: Nielsen MegaView Search

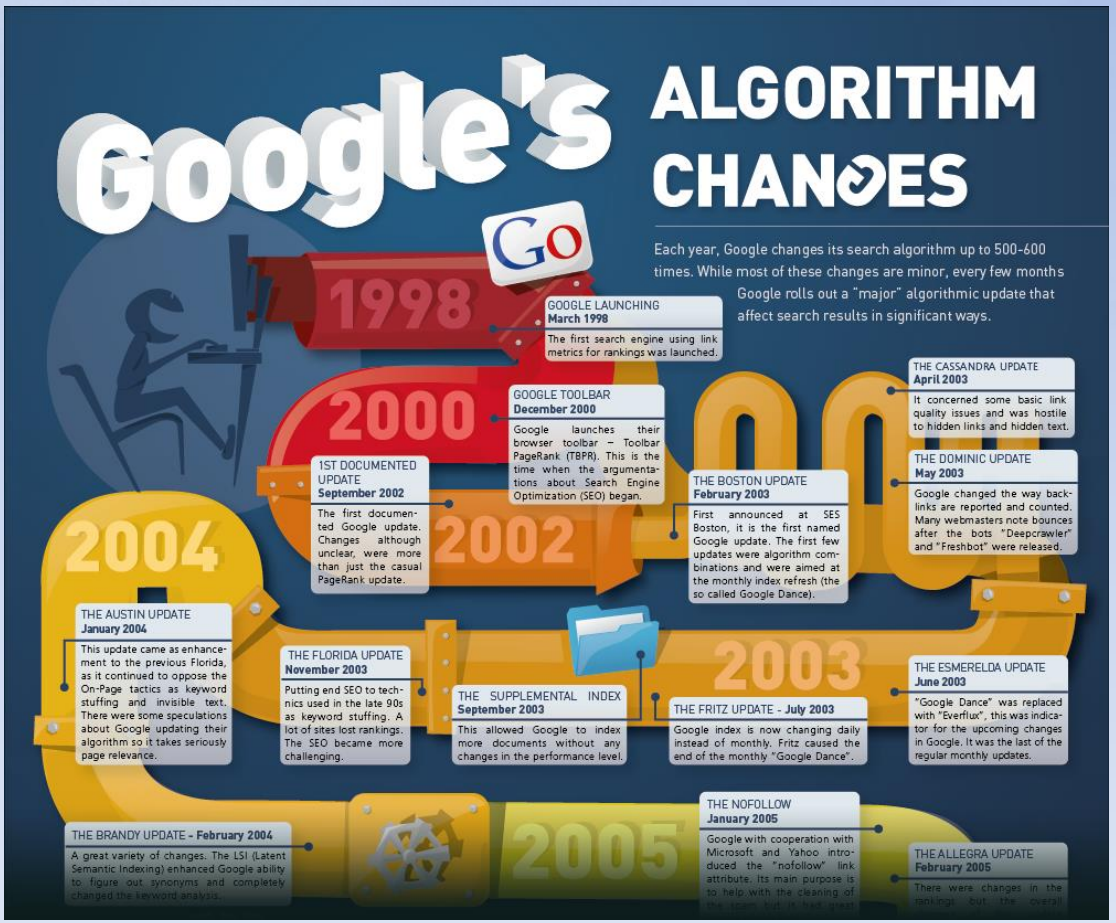
Popularność wyszukiwarek (3)

Worldwide Search Market Overview, July 2009 vs. July 2008				
	Searches (Millions)			
	July 2008	July 2009	Change (%)	Share (%)
Total Internet	*80,554*	*113,685*	*41*	*100*
Google sites	48,666	76,684	58	67.5
Yahoo! Sites	8,689	8,898	2	7.8
Baidu.com Inc.	7,413	7,976	8	7
Microsoft Sites	2,349	3,317	41	2.9
eBay	1,223	1,723	41	1.5
NHN Corporation	1,243	1,526	23	1.3
Ask Network	929	1,291	39	1.1
Yandex	663	1,290	94	1.1
AOL LLC	1,148	1,023	-11	0.9
Facebook.com	743	879	18	0.7

Źródło: comScore qSearch, 2009

Zmiany w Google'u (1998 – 2012)

http://www.seopalbg.com/blog/google-algorithm-changes-1998-2012-infographic



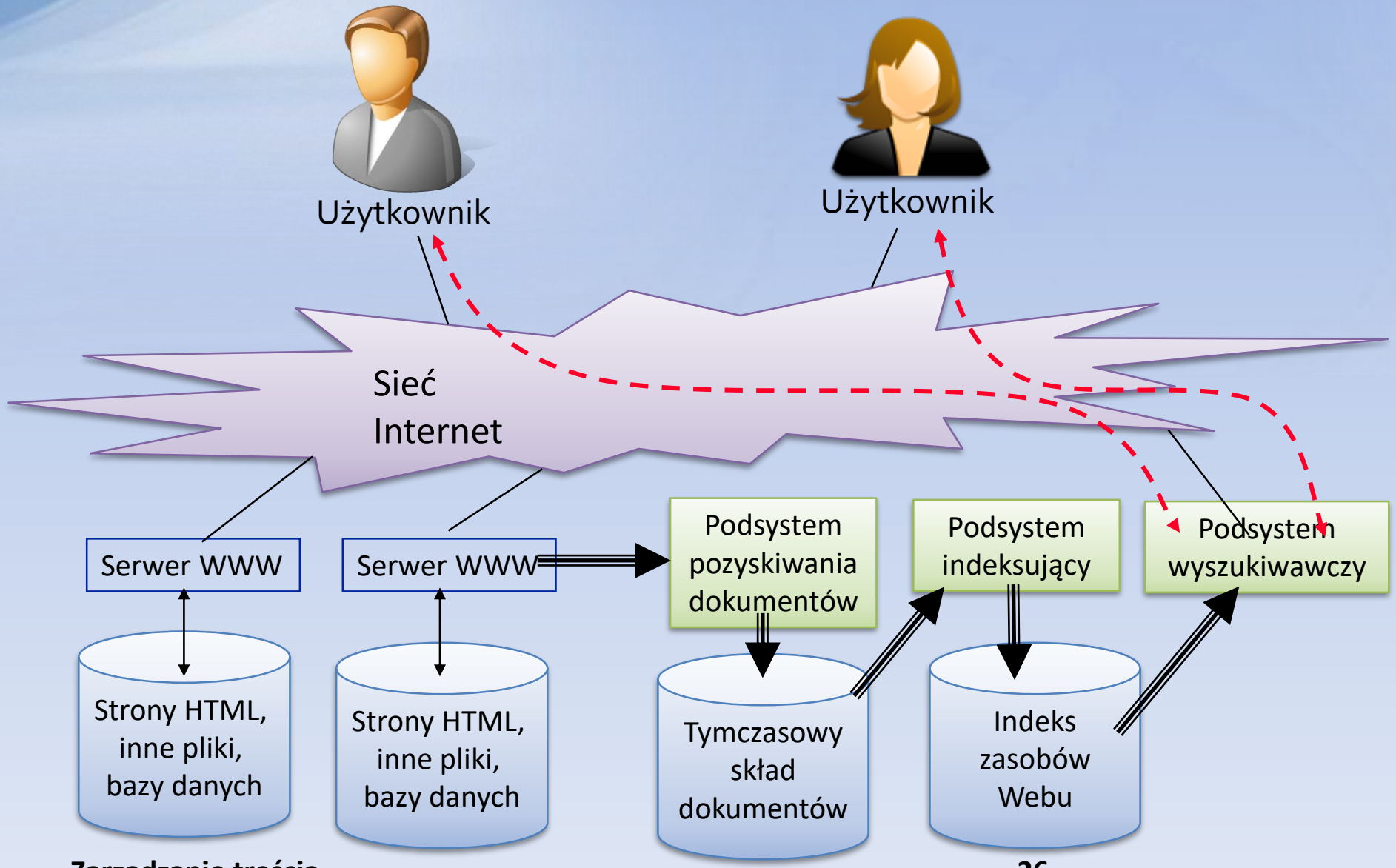
Paradygmaty wyszukiwania w Internecie

- Najbardziej interesujące jest korzystanie z uniwersalnej wyszukiwarki.
- Inne sposoby:
  - Katalogi stron, wyspecjalizowane katalogi przedmiotowe.
  - Własne zakładki i zestawienia linków tworzone przez użytkownika dla celów własnych; często udostępnione publicznie
  - Różnorodne FAQ (Frequently Asked Queries)
  - Zapytania posyłane na listy dyskusyjne news

Paradygmaty wyszukiwania w Internecie (2)

- Inne sposoby – c.d.:
  - Docieranie do stron domowych konkretnych użytkowników Internetu i przeglądanie informacji na tych stronach
  - Zapytania/prośby wysyłane przez email do konkretnych osób
  - Portale, wortale, strony startowe, wyszukiwarki w obrębie portali
  - Osobiste narzędzia wyszukiwawcze
  - Korzystanie z (płatnych) usług wyspecjalizowanych firm

Schemat systemu wyszukiwawczego



Podsystemy wyszukiwarki

- Podsystem pozyskiwania dokumentów – tzw. robot, pająk (spider) lub „crawler”. Obiega sieć w cyklu (np. co tydzień) ściągając na serwer dostępne dokumenty. Zasada ‘tranzytywnego domknięcia”: jeżeli ściągnie dokument X, to następnie ściąga wszystkie dokumenty, których URL-e znajdują się wewnątrz dokumentu X, z pominięciem już ściągniętych dokumentów.
- Podsystem indeksujący: po zebraniu porcji dokumentów następuje wybranie z dokumentu znaczących słów i wstawienie ich do centralnego indeksu, razem z odpowiednim URL-em i ewentualnie kontekstem.
  - Proces indeksowania może być wspomagany przez ludzi (kategoryzacja).

Podsystemy wyszukiwarki (2)

- Podsystem wyszukiwawczy: reaguje na zapytania internauty/użytkownika i dokonuje ekstrakcji z indeksu w postaci strony/stron HTML, które przesyła do zadającego zapytanie.
  - Kolejność przesyłanych pozycji indeksu ma ogromne znaczenie.
  - Stosuje się specjalne metody (Google, metoda oparta na tzw.”hubs”), które wyliczają „trafność” (pertinency) pozycji indeksu i szeregują przesyłane pozycje w kolejności zmniejszajacej się trafności.
  - Stosowany jest także klucz komercyjny (link do sponsora wyszukiwarki na początku przesyłanego zestawienia linków).

Szum, straty informacji, relewancja, trafność

- Są to cechy mierzalne liczbowo określające jakość rezultatu wyszukiwania.
- Szum informacyjny: informacja niechciana, zbędna, dostarczona wskutek mało precyzyjnego zapytania i/lub mało precyzyjnego mechanizmu wyszukiwawczego.
- Strata informacyjna: informacja pożądana, która nie została dostarczona wskutek mało precyzyjnego zapytania i/lub mało precyzyjnego mechanizmu wyszukiwawczego.

Szum, straty informacji, relewancja, trafność (2)

- Relewancja (*relevancy*): określa stopień w jakim wyszukiwane informacje formalnie pasują do zapytania.
- Trafność (*pertinency*): określa stopień, w jakim wyszukane informacje odpowiadają aktualnej potrzebie użytkownika.
  - Informacje mogą być relewantne, ale np. banalne, więc nietrafne.
  - Oczywiście trafność jest dużo ważniejsza od relewancji.

Język wyszukiwawczy użytkownika

- Lepiej byłoby nazwać to „metafora wyszukiwawcza”, bo coraz częściej nie są to języki, ale metafory graficzne, które sprzyjają naturalnemu zachowaniu sie użytkownika podczas wyszukiwania.
- Sformalizowany język wyszukiwawczy jest mało przyjemny dla użytkownika. Im bardziej sformalizowany, tym bardziej nieprzyjazny. Mocniej sformalizowany język nie musi oznaczać zwiększenia trafności. Tylko trafność ma znaczenie dla użytkownika.
- Brak sformalizowanego języka wyszukiwawczego jest mało przyjemne dla użytkownika, ponieważ sprzyja dostarczaniu szumu informacyjnego oraz obniża poziom trafności.

Język wyszukiwawczy użytkownika (2)

- Język musi odpowiadać kryteriom efektywności inżynierskiej:
  - Czas wyszukiwania, zużycie zasobów (np. pamięci)
  - Czas i koszt stworzenia całości systemu, koszt eksploatacji
  - Jakość usługi pozwalająca na zysk (zwykle finansowy)
- Czy jest tu złoty środek?

Inteligentne wyszukiwarki internetowe

- Marzenie i przedmiot działalności wielu ośrodków akademickich.
  - Czy wobec skali Webu i niezbędnych inżynierskich kompromisów nie jest to zbyt śmiałe marzenie?
- Web jest bardzo szczególną bazą danych
  - Obiekty mają wysoce niejednorodną strukturę, która nie jest określona; jeżeli nawet jest określona, to obiekty mogą mieć błędną strukturę.

Inteligentne wyszukiwarki internetowe (2)

- Web jest bardzo szczególną bazą danych – c.d.
  - Liczba obiektów i ich rodzajów stale rośnie
  - Obiekty są zróżnicowane tematycznie i znaczeniowe, informacje mogą być formalnie i merytorycznie niespójne.
  - Obiekty tworzą szczególną sieć semantyczną poprzez hiper-linki. Semantyczne znaczenie hiper-linków może być dowolne.
  - Znaczenie obiektu może być wyznaczone przez związane z nim obiekty

Inteligentne wyszukiwarki internetowe (3)

- Jednocześnie potencjalny użytkownik korzystający z Webu jest bardzo wymagający:
  - Nie jest i nie chce być informatykiem; informatyczny żargon odbiera jako nieprzyjazny bełkot.
  - Nienawidzi czytania jakichkolwiek instrukcji obsługi, „pomocy” i porad.
  - Szybko opanowuje pewne wzorce postępowania przy pracy z Internetem i bardzo niechętnie je zmienia lub modyfikuje.

Inteligentne wyszukiwarki internetowe (4)

- Potencjalny użytkownik – c.d.
  - Zna prawie doskonale (z pozycji informatyka) swoją dziedzinę działalności zawodowej i oczekuje profesjonalnych odpowiedzi na swoje zapytania,
  - ... ale nie przywiązuje wagi do precyzyjnego, formalnego zadawania pytań.
  - Oczekuje pomocy w przypadku jakichkolwiek trudności oraz przyjacielskiego zachowania się komputera w każdej sytuacji.
  - Nie starcza mu cierpliwości do oglądania setek dokumentów, zwykle traci cierpliwość po 10-tym nietrafnym dokumencie.
  - Zależy mu na szybkim uzyskaniu adekwatnych wyników.

Inteligentne wyszukiwarki internetowe (5)

- Niektórzy uważają, że oznacza to konieczność wbudowania do wyszukiwarek pewnej inteligencji. Może to odbywać się np. poprzez następujące mechanizmy:
  - Informowanie użytkowników o istotności dokumentu, np. prezentacja automatycznie generowanych „streszczeń”, innych słów kluczowych, wag ważności dokumentu.
  - Adaptację (poprzez automatyczne „uczenie się” systemu, personalizację) do indywidualnych preferencji użytkownika.
  - Podpowiedzi co do dalszych lub alternatywnych kierunków poszukiwań.
  - Uwzględnienie różnorodnych statystyk i pomiarów dotyczących zachowania się całej populacji użytkowników, celem odkrycia pewnych prawidłowości.

Inteligentne wyszukiwarki internetowe (6)

- Uważa się, że to wymaga automatycznego „rozumienia” (?) tekstu, oraz zastosowania technologii „odkrywania wiedzy”: m.in. algorytmów klasyfikacji i grupowania informacji.
- Aktualnie nie udało się zastosować tego typu rozwiązań na masową (komercyjną) skalę.

Wyszukiwanie zorientowane geograficznie

- Wyszukiwarka dodatkowo przechowuje informacje o geograficznej lokalizacji informacji, np. Polska czy nawet Warszawa,
- Znając lokalizację użytkownika (np. Polska) można zwiększyć wagę rezultatów pochodzących z serwisów z tej samej (lub zbliżonej) lokalizacji,
- Dodatkowo można też uwzględniać natywny język użytkownika,
- Umożliwia znaczące polepszenie jakości rezultatów
- Daje szanse realizacji nowej kategorii usług, np. znajdź najbliższą pizzerię.

Wyszukiwanie przy użyciu języka naturalnego

- Rodzi ogromne problemy, szczególnie w językach z bogatą morfologią, takich jak polski lub niemiecki.
  - Niektóre z tych problemów są znacznie zredukowane w języku angielskim.
- Zaletą języka naturalnego jest to, że użytkownik go zna (a przynajmniej tak mu się wydaje) i nie musi się go specjalnie uczyć.
- Jest on również elastyczny, pozwala wyrazić dowolną informację.

Wyszukiwanie przy użyciu języka naturalnego (2)

- Są jednak liczne wady języka naturalnego jako środka wyszukiwania :
  - Jest nieformalny i nieformalizowalny (szczególnie semantyka) na obecnym etapie wiedzy. W związku z tym automatyczne „rozumienie” tekstu jest na dzisiaj wyłącznie pseudo-naukową retoryką (antropomorfizmem).
  - Ta sama informacja może być wyrażona na dowolną ilość sposobów, co powoduje trudności przy automatycznym określaniu zgodności.
  - Informacja może być różnie rozumiana przez różne osoby.
  - Te same wyrazy lub zdania mogą mieć różne znaczenie zależnie od dowolnie rozległego kontekstu i skojarzeń powstających w umyśle odbiorcy.

And, Or, Not

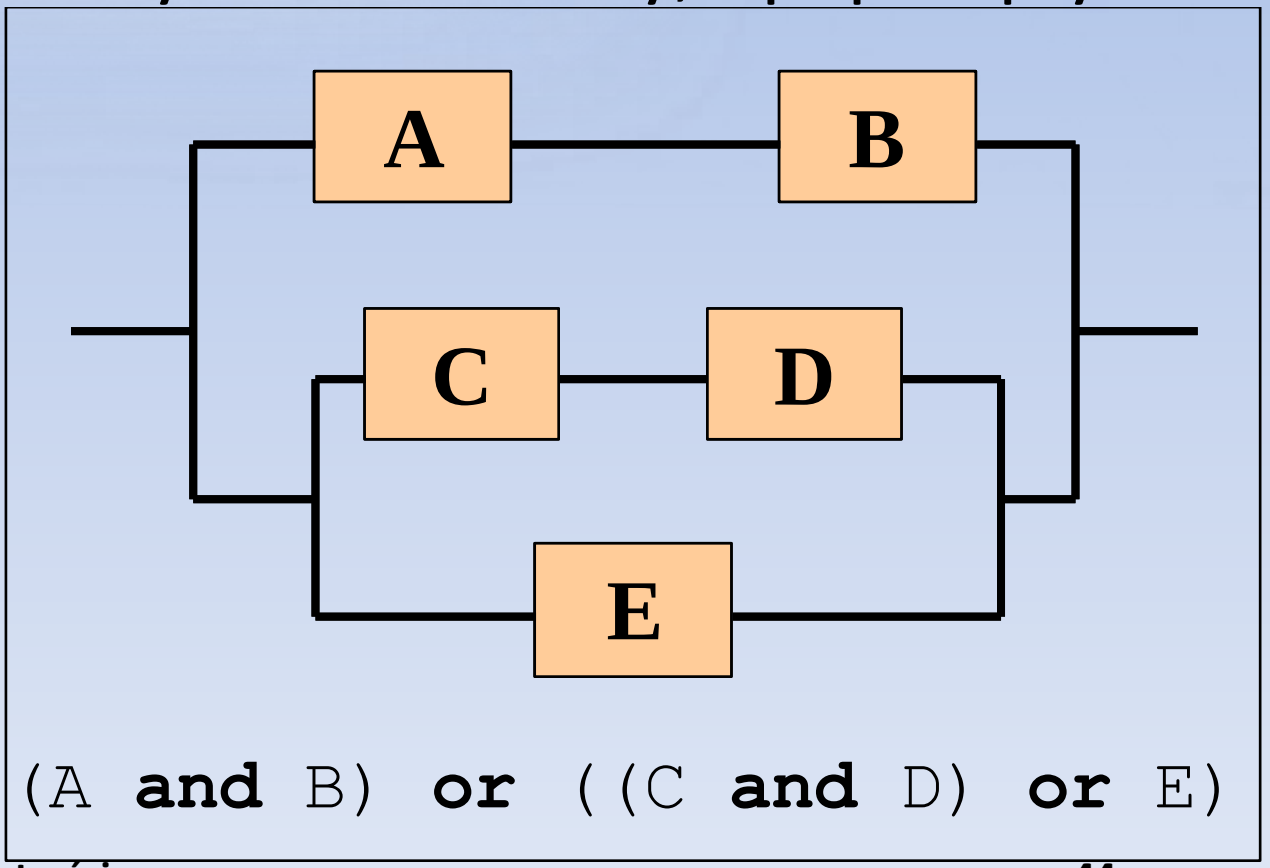
- Wyszukiwanie poprzez prostą algebrę Boola. Patrz Google Advanced Search.
  - Indeks dokumentów jest zbiorem, gdzie każdy element (pozycja) jest opatrzony zbiorem słów kluczowych
  - Czyli dla każdego słowa kluczowego mamy podzbiór pozycji indeksu – tych, które są opatrzone tym słowem.
  - Możemy tworzyć wyrażenia złożone ze słów kluczowych, operatorów AND, OR, NOT i nawiasów.
  - Operator AND działający na dwóch podwyrażeniach oznacza teorio-mnogościowe przecięcie odpowiednich podzbiorów pozycji indeksu.
  - Operator OR działający na dwóch podwyrażeniach oznacza teorio-mnogościową sumę odpowiednich podzbiorów pozycji indeksu.
  - Operator NOT działający na podwyrażeniu oznacza zbiór powstały poprzez odjęcie z całości indeksu pozycji wyznaczonych przez podwyrażenie.

Problemy z naiwnym wyszukiwaniem poprzez And/Or/Not

- Metoda nie uwzględnia syntagmatyki, czyli sąsiedztwa wyrazów w tekście. Nie jest obojętne, czy wyrazy łączone przez AND są obok siebie, czy też są odległe o kilka stron.
- Metoda nie uwzględnia fraz, czyli sekwencji wyrazów. Czasami trzeba dokładnie szukać całej frazy, a czasami pojedynczego wyrazu z tej frazy.
- Metoda nie uwzględnia fleksji, czyli odmian wyrazów poprzez przypadki, rodzaje, czasy, liczbę, itd.
- Metoda nie uwzględnia paradygmatyki, czyli semantyki wyrazów, w szczególności zawierania się zakresów znaczeniowych.
- Metoda nie uwzględnia synonimii (różnych wyrazów dla określenia tego samego) i homonimii (identycznego wyrazu dla określenia różnych rzeczy)
- Metoda nie uwzględnia popularnych błędów pisowni (np. braku ogonków, błędów ortograficznych, itd.).

Problemy z naiwnym wyszukiwaniem poprzez And/Or/Not (2)

- Użytkownicy mylą znaczenie AND/OR
- Alternatywne metafory, np. przepływowa



Syntagmatyka

- Związek z ciągiem tekstowym, związek składniowy pomiędzy wyrazami w tekście.
  - jeden kawałek tekstu „...żołnierz strzela...”
  - drugi kawałek tekstu „...dziewczyna mruga”
  - zapytanie „dziewczyna strzela” dostarczy błędnie ten dokument ponieważ nie uwzględniono związku syntagmatycznego.
- Uwzględnienie syntagmatyki oznacza konieczność zaindeksowania kontekstowego, gdzie pozycja indeksu będzie określona poprzez zestaw wyrazów lub fraz, które ze sobą sąsiadują.
- Zwiększy to precyzję wyszukiwania, ale musi spełniać jeszcze kryteria inżynierskiej efektywności.
- W języku polskim kolejność wyrazów, jak również ich odległość w tekście nie jest często wyróżnikiem, dlatego potrzebne są proste heurystyczne kryteria dla określenia związku składniowego.
- W Google związek składniowy wpływa na ocenę trafności.

Frazy

- Dość często sekwencje wyrazów posiadają określone znaczenie, specyficzne dla tej sekwencji.
- Użytkownik może poszukiwać informacji na podstawie często zasłyszanej sekwencji wyrazów.
- W języku polskim dodatkowym utrudnieniem jest fakt, że porządek wyrazów w takiej frazie może być zmieniany:
  - „zmaterializowana perspektywa” <==> „perspektywa zmaterializowana”

Frazy (2)

- Konsekwencją jest konieczność wprowadzenia do indeksu nie tylko pojedynczych wyrazów, ale także ich często spotykanych sekwencji.
  - Problemem jest jak identyfikować takie frazy, jak zorganizować automatycznie indeks uwzględniający takie frazy, i jak używać tego indeksu
  - Można znaleźć pewne heurystyczne reguły pozwalające traktować sekwencje wyrazów jako pojedynczy element wyszukiwawczy.
  - W takim przypadku zarówno fraza, jak i jej składowe stanowią pozycje indeksu.

Fleksja

- Oznacza odmiany wyrazu:
  - „kot”, „koty”, kotu”, „kotem”, „kotami”,...
  - „zielony”, „zielono”, „zazielenić”, „zazieleniony”, „zielone”, „zieleń”,...
- Jest to jak dotąd najpoważniejszy problem przy budowie wyszukiwarek, szczególnie w języku polskim, gdzie fleksja jest bardzo bogata.
  - Dla niektórych wyrazów doliczono się ponad 100 form fleksyjnych
- W wyszukiwarkach angielskojęzycznych problem jest mniejszy i w wielu przypadkach można go sprowadzić przy indeksowaniu dokumentów i wstępnym przetwarzaniu zapytań do obcięcia końcowego –s lub –es.

Fleksja (2)

- W języku polskim zastosowanie podobnej metody, polegającej na obcięciu kilku znaków z końca i/lub z początku przy pomocy prostych reguł formalnych prowadzi do znacznego (nieakceptowalnego) szumu informacyjnego,
  - który w anegdotyczny sposób ośmieszy naszą wyszukiwarkę w oczach użytkowników.
- Częściowo można rozwiązać ten problem przy pomocy specjalnych słowników. Niestety, mają one b. dużą objętość.

Paradygmatyka

- Oznacza semantyczne zależności pomiędzy wyrazami lub frazami, które są niezależne od ich użycia w tekście.
  - Najbardziej popularnym tego rodzaju związkiem jest zawieranie się zakresów znaczeniowych wyrazów, np. „maszyna rolnicza” ← „traktor”. W indeksie występuje „traktor”, a w pytaniu użytkownika jest „maszyna rolnicza”. Jak nasza wyszukiwarka skojarzy te pojęcia?
  - Drugim istotnym tego rodzaju związkiem jest instancjacja, czyli związek pomiędzy pojęciem i konkretnym obiektem, który to pojęcie oznacza: np. „prezydent” ← „Lech Kaczyński”.

Paradygmatyka (2)

- C. d.
  - Trzecim istotnym związkiem jest synonimia, do której zaliczamy także skróty, skrótowce i akronimy, np. „XML” 🡐🡒 „extended markup language”, „traktor” 🡐🡒 „ciągnik”, „database view” 🡐🡒 „view”, ...
  - Istnieje wiele dalszych tego rodzaju związków, w szczególności związek część-całość np. „samolot” 🡐 „śmigło”, i dowolny związek skojarzeniowy (patrz też), np. „Jurek Owsiak” patrz też „działalność charytatywna”.
- Bardzo kosztownym sposobem opanowania paradygmatyki jest ręczna budowa odpowiednio zorganizowanych słowników (tzw. tezaurusów).

Popularne błędy pisowni

- Zarówno twórcy tekstów, jak i użytkownicy wyszukiwarek są omylni.
  - Ten fakt przez długie lata nie docierał do twórców systemów wyszukiwania informacji, których modele były idealistyczne – zakładały bezbłędność indeksowanych tekstów i bezbłędność zapytań.
  - W systemach, gdzie użytkownik stanowi jedyne kryterium powodzenia, nie uwzględnienie jego potencjalnych błędów jest błędem biznesowym.

Popularne błędy pisowni (2)

- Błędy ortograficzne, gramatyczne charakterystyczna dla konkretnego języka.
- Są również inne popularne błędy, takie jak czeski błąd (zwany w Czechach „polskim błędem”), pomijanie liter, itd.

Problem inżynierskiej/biznesowej efektywności

- Świat akademicki ma tendencję do idealistycznego traktowania poprzednio wymienionych problemów.
- Metoda „brute force” (brutalna siła): rozwiązać problem tak, jak on występuje, w izolacji od innych problemów.
  - Np. wiele ośrodków zabrało się niegdyś za budowę własnych tezaurusów dla opanowania problemu paradygmatyki; następnie po paru latach tezaurusy te zasiliły składy makulatury z powodu dezaktualizacji.

Problem inżynierskiej/biznesowej efektywności (2)

- To spowodowało, że świat komercyjny podchodzi z nieufnością do rezultatów produkowanych przez świat akademicki.
  - W praktyce, ignoruje te rezultaty i wynajduje własne rozwiązania.
  - Rozwiązania te stawiają często pod wielkim znakiem zapytania w/w klasyczne pojęcia w zakresie wyszukiwania informacji.
- Inżynierska/biznesowa efektywność jest wyznaczona przez zadowolenie użytkownika i związane z tym powodzenie przedsięwzięcia.
  - Inne kryteria są drugorzędne.

Rozwiązania usprawniające wyszukiwanie

- Zapamiętywanie poprzednich zainteresowań (kryteriów wyszukiwania) użytkownika.
- Śledzenie odwiedzanych stron Web'u (wymaga zainstalowania dedykowanego oprogramowania klienckiego),
- Połączenie wyszukiwania w Internecie z przeszukiwaniem lokalnego komputera (MS Windows Desktop Search, Google Desktop Search)
- Przyporządkowanie wag do poszczególnych słów; popularne wyrazy mają niskie współczynniki
- *Więcej przy okazji omawiania Google'a*

Prywatność użytkowników

- Wiele narzędzi wyszukujących gromadzi różne informacje o użytkownikach
- W niektórych przypadkach może to naruszać ich prywatność
- Rozwiązania chroniące prywatność:
  - Serwisy internetowe, np. http://www.megaproxy.com/
  - Oprogramowanie klienckie, np. http://www.anonymizer.com
- Uzyskanie „złotego środka” nie jest łatwe

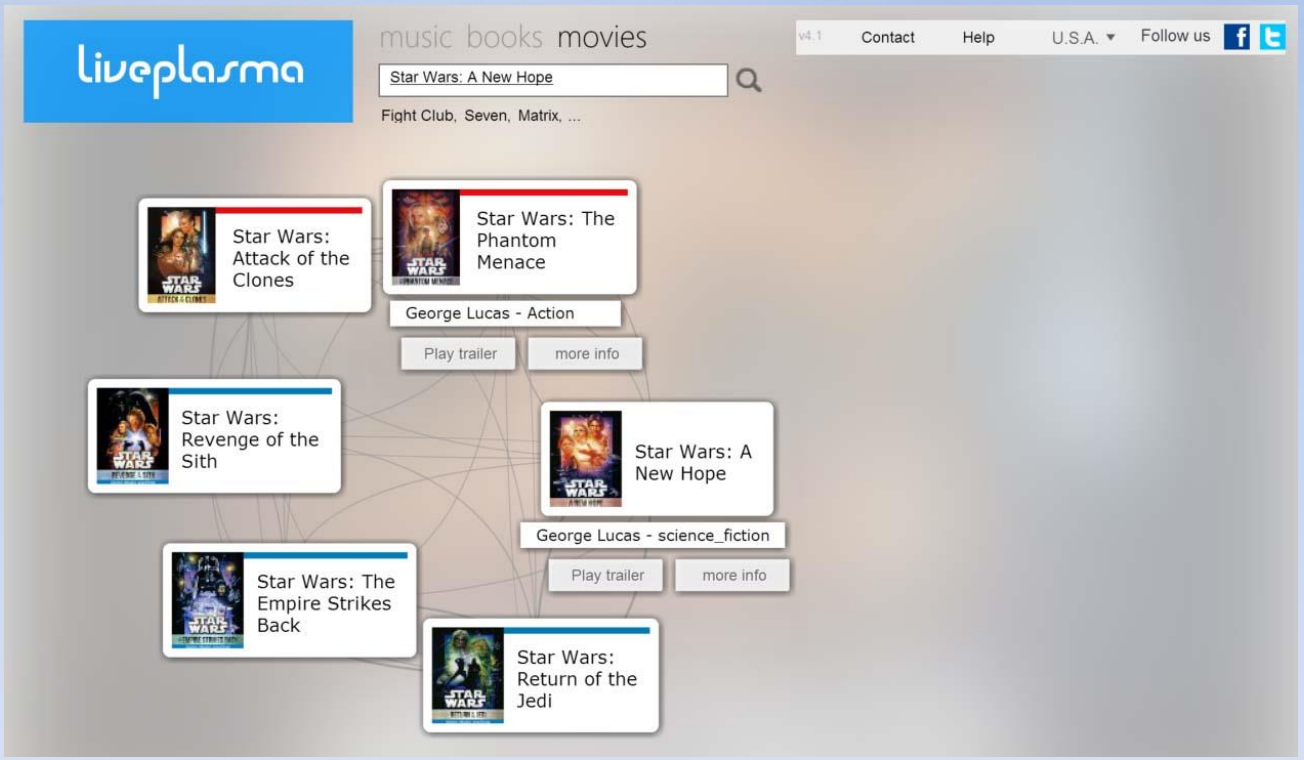
Nowe rozwiązania

- Eyeplorer (http://en.eyeplorer.com/)

Wyniki dla frazy: ***Arduino***

Nowe rozwiązania (2)

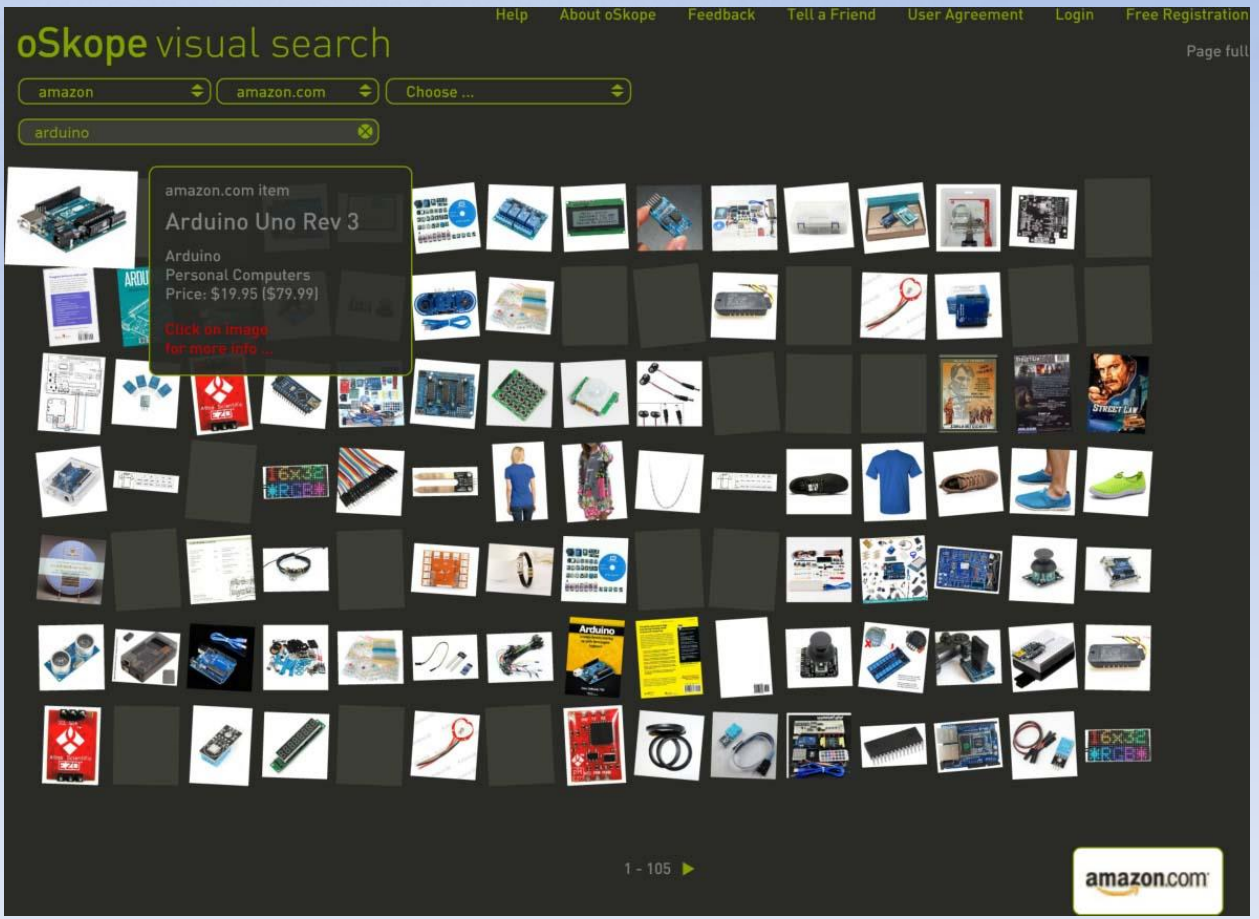
- LivePlasma (http://www.liveplasma.com/)



Wyniki dla frazy: ***Star Wars: A New Hope***

Nowe rozwiązania (3)

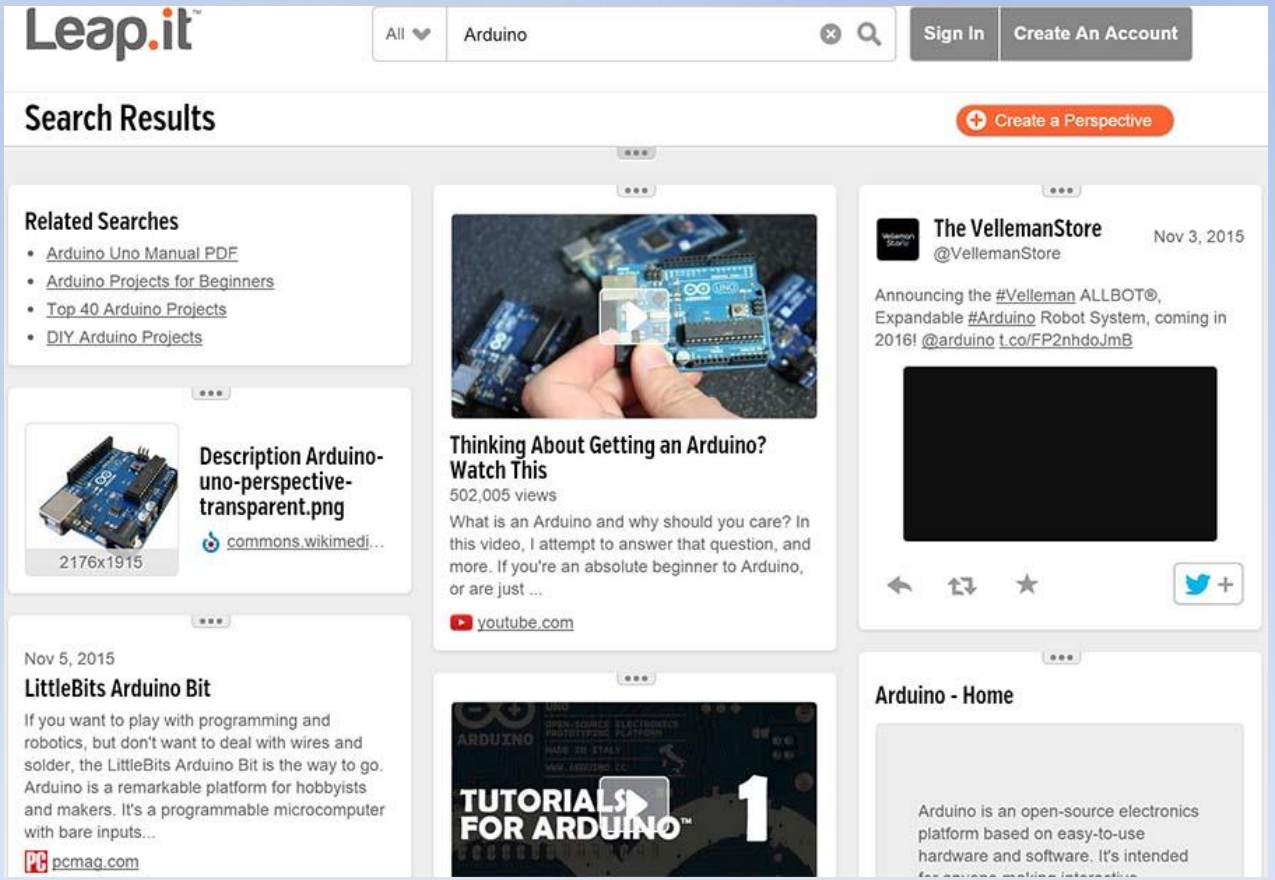
- oSkope (http://www.oskope.com/)



Wyniki dla frazy: ***Arduino***

Nowe rozwiązania (4)

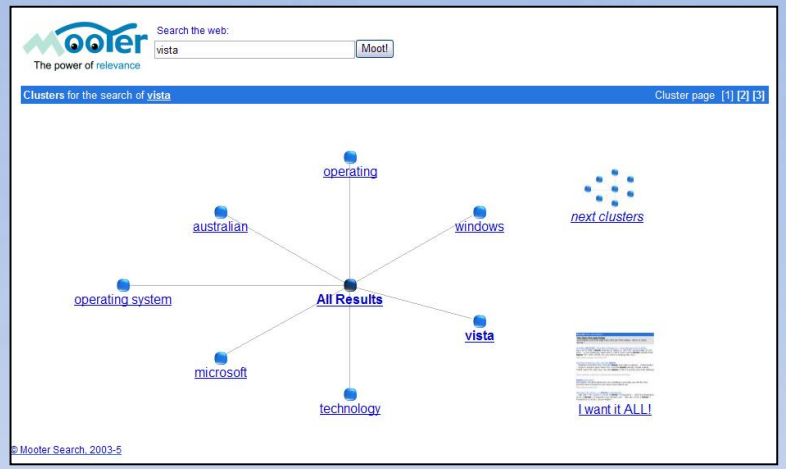
- Leap It (http://leap.it/)



Wyniki dla frazy: ***Arduino***

Nowe rozwiązania (5)

- Serwisy, które zakończyły działalność:
  - Mooter – skupiska (clusters) (http://www.mooter.com/)
  - Kartoo (http://www.kartoo.com/)



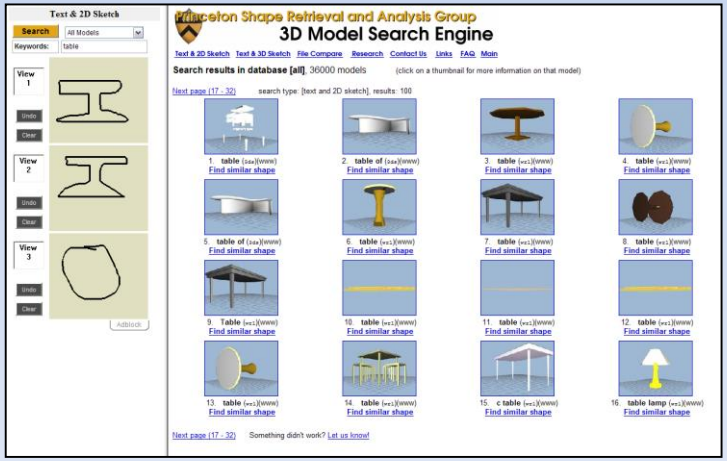
Wyszukiwanie innych mediów

- Wyszukiwanie za pomocą innych paradygmatów niż tekstowe (np. grafiki czy muzyki), nastręcza wiele trudności.
- Princeton 3D Model Search Engine umożliwia wyszukiwanie modeli 2D oraz 3D (http://shape.cs.princeton.edu/search.html)
  - Słowa kluczowe
  - Rozpoznawanie kształtów na podstawie szkiców tworzonych przez użytkownika
  - Efekty są obiecujące,

    ale wydaje się, że

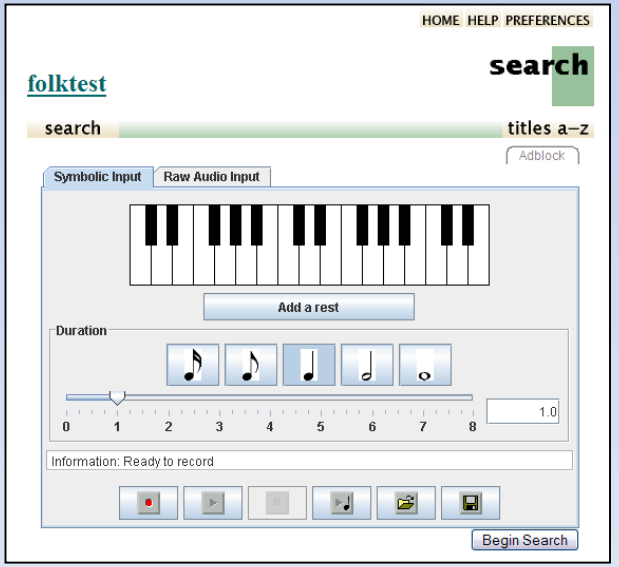
    wymaga jeszcze

    dopracowania



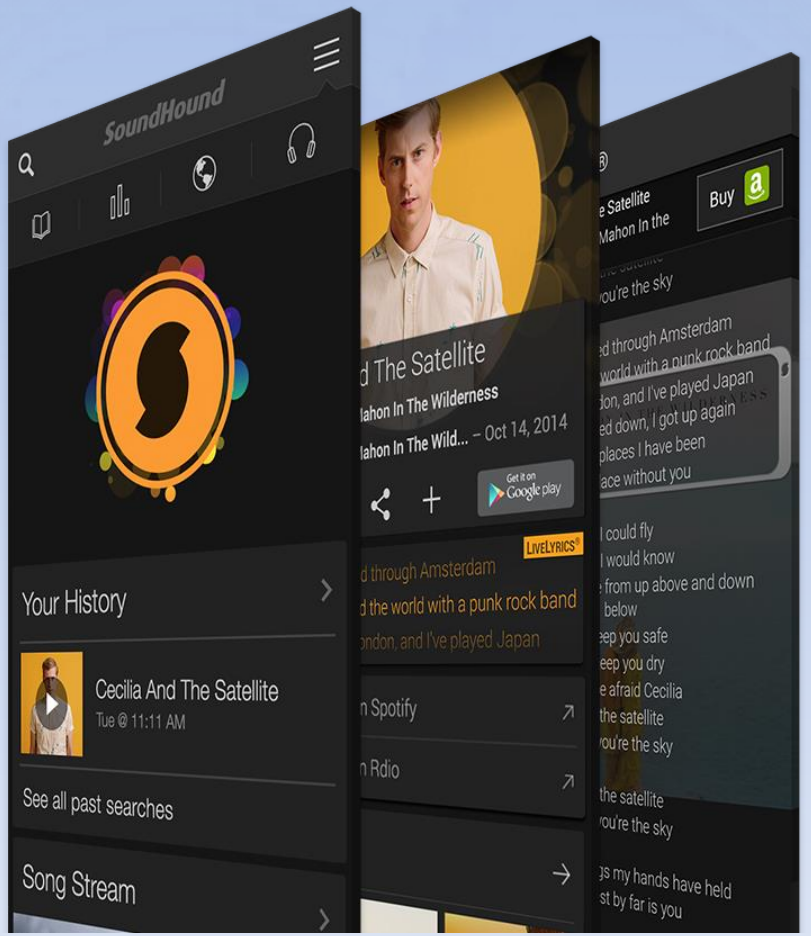
Wyszukiwanie innych mediów (2)

- Wyszukiwanie muzyki
  - Jak zdefiniować zapytanie?
    - Notacja muzyczna (np. nuty) lub jakiś jej wariant
    - Odegranie kawałka utworu korzystając z klawiatury
    - Zanucenie fragmentu do mikrofonu
    - Tekst piosenki
  - Meldex stworzony w ramach New Zealand Digital Library Project (http://www.nzdl.org/musiclib)



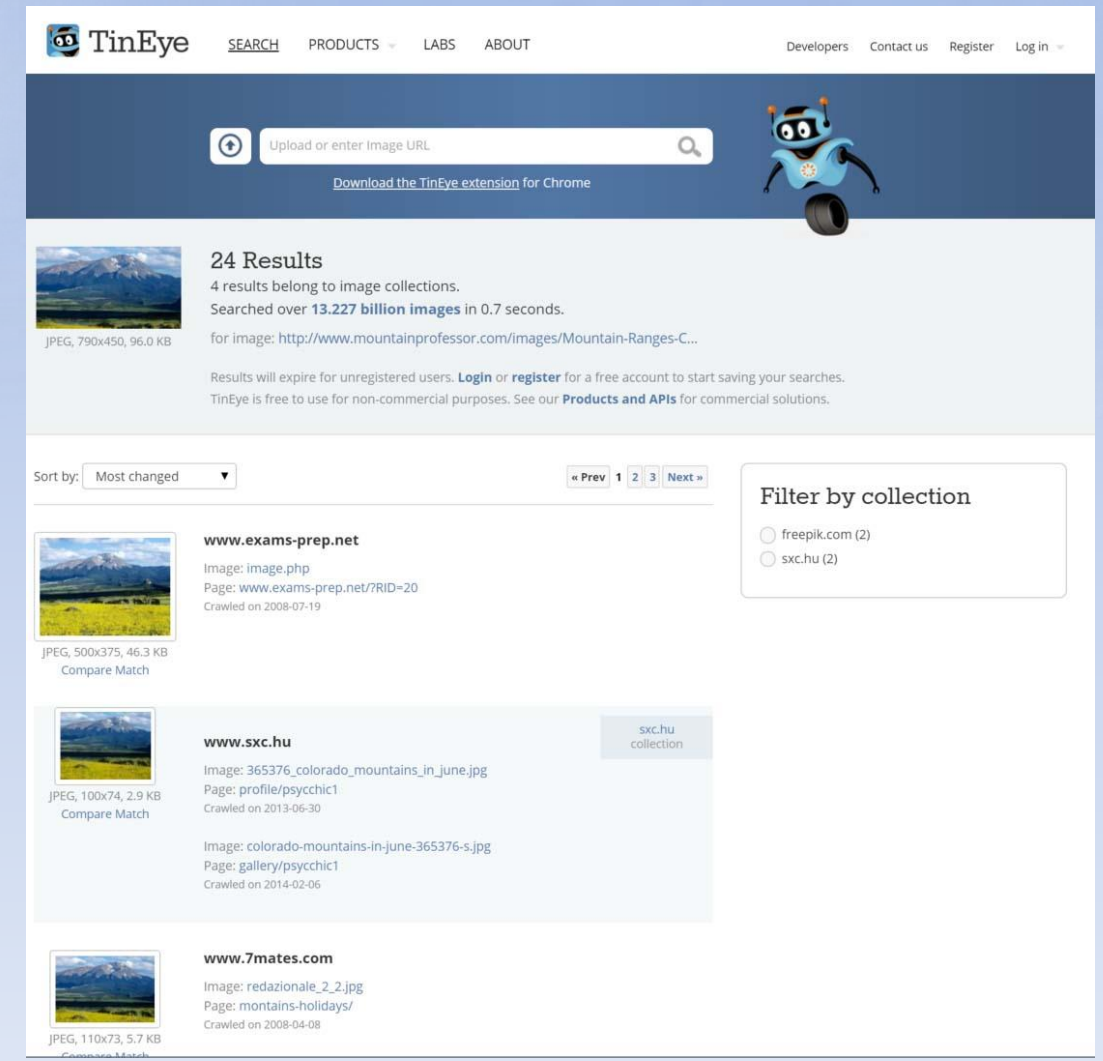
Wyszukiwanie innych mediów (3)

- Wyszukiwanie muzyki – c. d.
  - Aplikacja mobilna **SoundHound** (http://www.soundhound.com/) automatycznie rozpoznaje odgrywaną muzykę.



Wyszukiwanie innych mediów (4)

- Wyszukiwanie grafiki
  - **TinEye** (http://tineye.com/) rozpoznaje duplikaty grafiki (nawet zmodyfikowane)
  - Funkcja w Google Chrome



Podsumowanie

- Wyszukiwanie informacji w Internecie jest problemem bardzo złożonym.
- Wyszukiwarki, nie tylko ułatwiają korzystanie z Internetu, ale wręcz je umożliwiają.
- Obecnie rynek jest zdominowany przez jedną firmę: Google.
- Będzie tak do czasu, aż ktoś inny wymyśli, nowe, przełomowe sposoby wyszukiwania.
- Jak dotąd nie ma dobrych rozwiązań wyszukujących za pomocą innych metod niż tekstowe.
- Rezultaty zwracane przez popularne serwisy są coraz lepsze, ale wciąż można chcieć więcej.