7

Internetové **Technologie**

vyhledávání na internetu

Ing. Michal Radecký, Ph.D.
www.cs.vsb.cz/radecky

Vyhledávání a vyhledávače

- Jediný možný způsob, jak získat obecný přístup k informacím na Internetu
- Nástroj (server, aplikace, apod.) nabízející služby pro vyhledávání požadovaných informací na základě specifikace zadání od uživatele. Toto vyhledávání se provádí nad daty, která jsou pro tento účel pořízena a udržována.

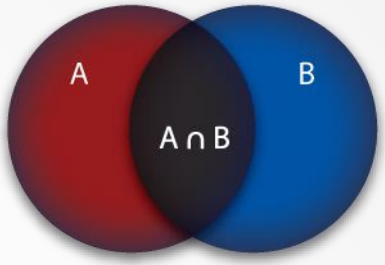
Vyhledávače

- Dělení podle architektury
  - centralizované (seznam.cz, Google, atd.)
  - decentralizované (Gnutella, FreeNet, atd.)
  - hybridní (Napster, BitTorrent, atd.)
- Dělení podle obsahu a služeb
  - katalog (firmy.cz, seznam.cz, centrum.cz, atd.)
  - fulltextový vyhledávač (seznam.cz, Google, atd.)
  - Sociální (Twitter, Facebook, atd.)
  - Sémantické

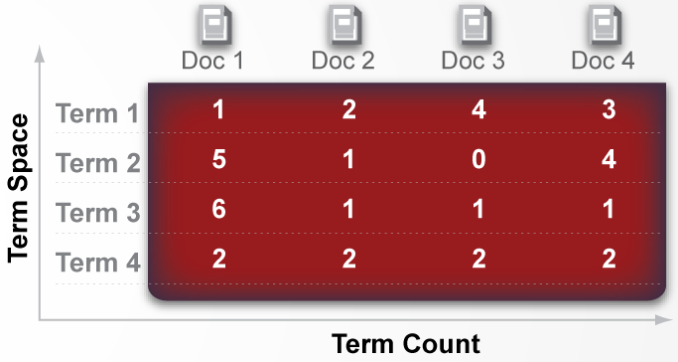
Vyhledávače

Zdroj: http://www.llrx.com/features/searchenterprise.htm

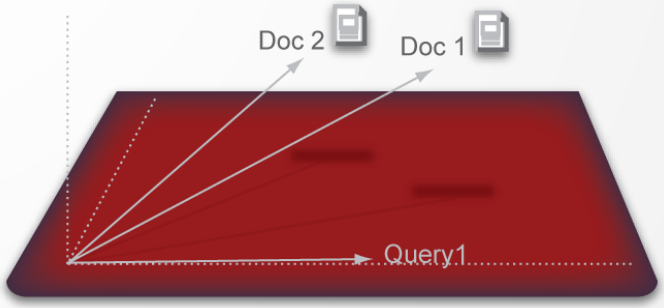
- Dělení podle modelu
  - Booleovský model (množiny)



  - Vektorový model



  - Fuzzy booleovský model



  - Shlukování, atd.

Vyhledávače

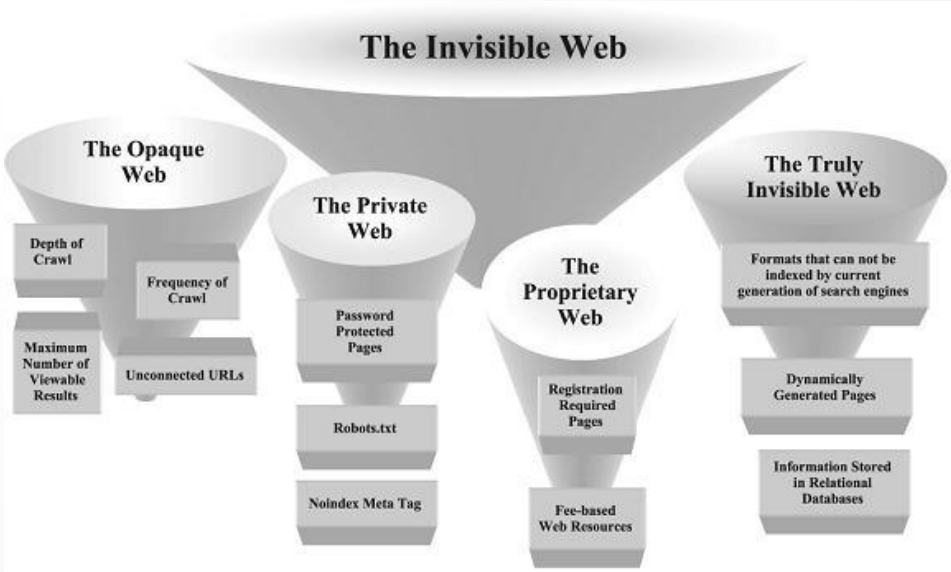
- Centralizované
  - jádro tvoří centralizovaná databáze (index) vytvářená pomocí „robotů“
  - architektura klient/server
  - problematické zajištění aktuálnosti databáze a tvorby indexů nad různými typy dat
  - rychlé vyhledávání relevantních informací
  - snadná správa fyzických dat
- Decentralizované
  - architektura peer-to-peer s využitím „floodingu“
  - aktuálnost hledaných dat odpovídá realitě
  - variabilita různých forem dotazů a nalezených dat
  - pomalá rychlost samotného vyhledávání a vysoké nároky na vytížení sítě
  - problematická správa dat z globálního pohledu
- Hybridní

Vyhledávače

- Katalogové
  - základ je databáze obsahující stromovou strukturu odkazů a informací o nich
  - plnění katalogů je především manuální
  - poskytují informace, kdy jejich relevantnost a aktuálnost závisí na aktualizaci informací o každé položce stromu zvlášť
  - dnes zpravidla propojené s fulltextovými vyhledávači
- Fulltextové
  - základ je rozsáhlá databáze (index) obsahující informace o stránkách a jejich obsahu
  - správa a údržba dat je automatizovaná, a to pomocí „robotů“
  - poskytují informace, kdy jejich relevantnost a aktuálnost závisí na periodicitě a možnostech „robota“ a indexování, využívá se ohodnocování jednotlivých položek
  - dnes se již možnosti indexace a vyhledávání rozšiřují i na dokumenty jiného formátu než WWW
- Sociální
  - v podstatě kombinace fultextového a katalogového vyhledávání
  - základem je zaměření na specifický typ obsahu a informací
- Sémantické

Typy vyhledávání

- Typy dotazů
  - Navigační dotazy
    - přístup na konkrétní stránku
    - „české aerolinie“ -> http://www.csa.cz
  - Informační dotazy
    - získání konkrétní informace
    - „počasí Praha“, „skoda fabia recenze”
  - Transakční dotazy
    - nalezení informace pro následnou akci
    - vyhledávání zboží, souborů, apod.
- Doménové oblasti
  - Obecné vyhledávání
  - Oborové vyhledávání
  - Vertikální vyhledávání
  - Vyhledávání v hlubokém webu (deep/invisible web)
  - Meta-vyhledávání
  - (www.qwiki.com)



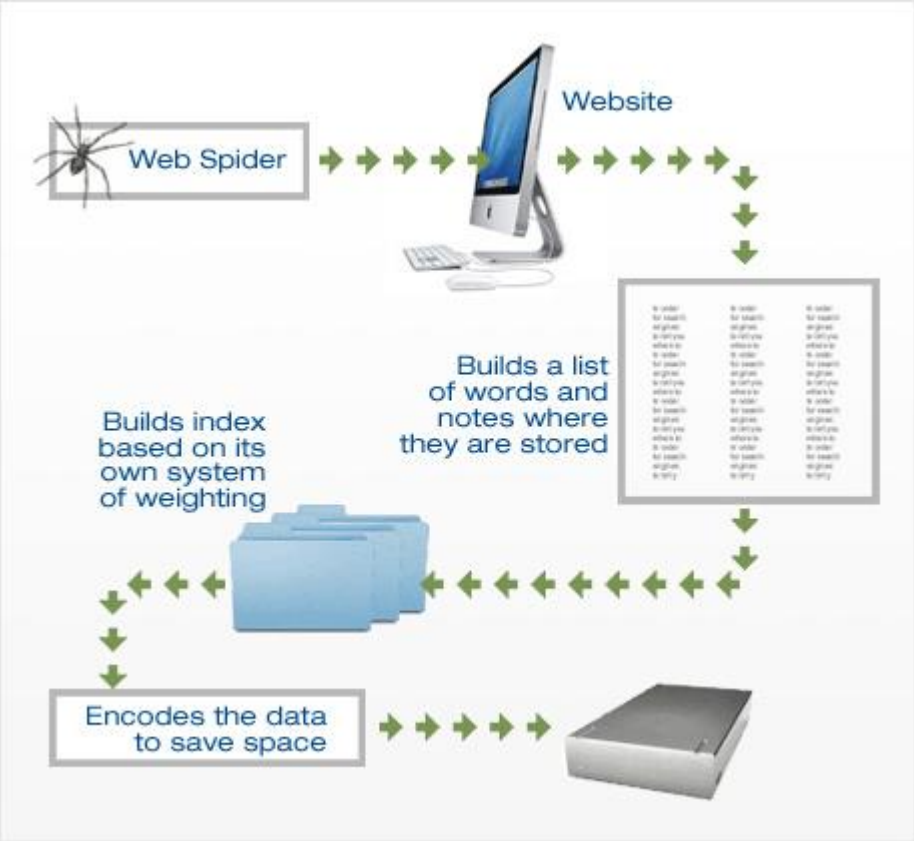
Source: Based on Sherman and Price (2001)

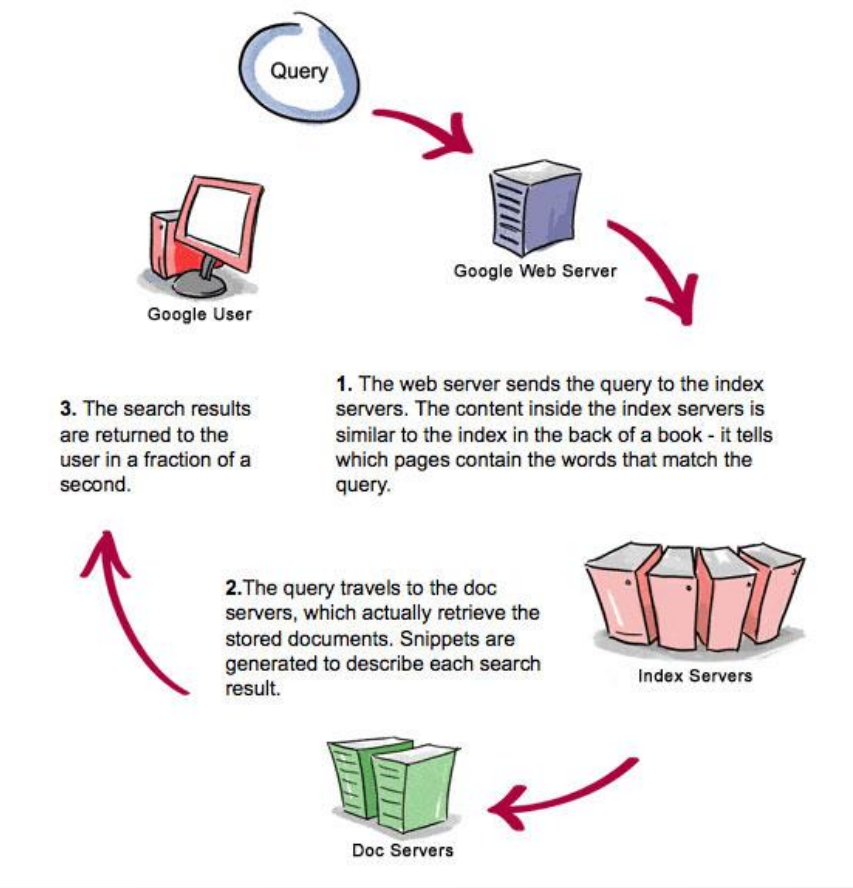
Fulltextové vyhledávače

- Fulltextové vyhledávání – technika pro hledání informací založena na zkoumání každého slova ve zdrojových datech (dokument, databáze, apod.)
- 3 fáze funkčnosti vyhledávače (Search Engine)
  - sběr dat - robot, spider, web crawler
  - uložení dat do databáze – index
  - dotazování
- Google.com, Yahoo.com, Altavista.com, Seznam.cz, Centrum.cz, atd.

Zdroj: hhttp://myblogranksfirstingoogle.blogspot.cz/2010/09/work-flow-of-search-engines.html

Fulltextové vyhledávače

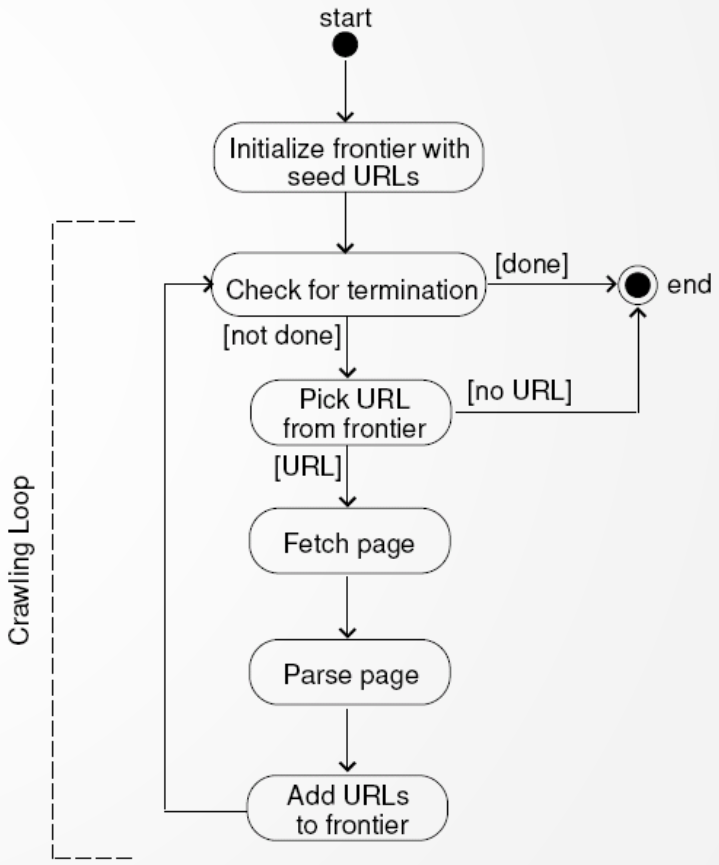




Crawler

Zdroj: http://dollar.biz.uiowa.edu/~pant/Papers/crawling.pdf

- Program, který po svém spuštění realizuje první fázi provozu vyhledávače
- Jedná se v podstatě o princip procházení grafu
- Vytvářejí kopie stránek v úložišti systému
- Zpracovávají data podle svého určení (obrázky, dokumenty, apod.)
- Zpravidla využívá parsování pouze na úrovni textových dat (HTML, XML, apod.)
- Obvykle pracuje s omezením počtu či hloubky zanoření
- Paměť pro již zpracované stránky
- Již dříve zpracované stránky se navštěvují znovu z důvodu nalezení změn
- Analyzují meta-tagy a soubor robots.txt



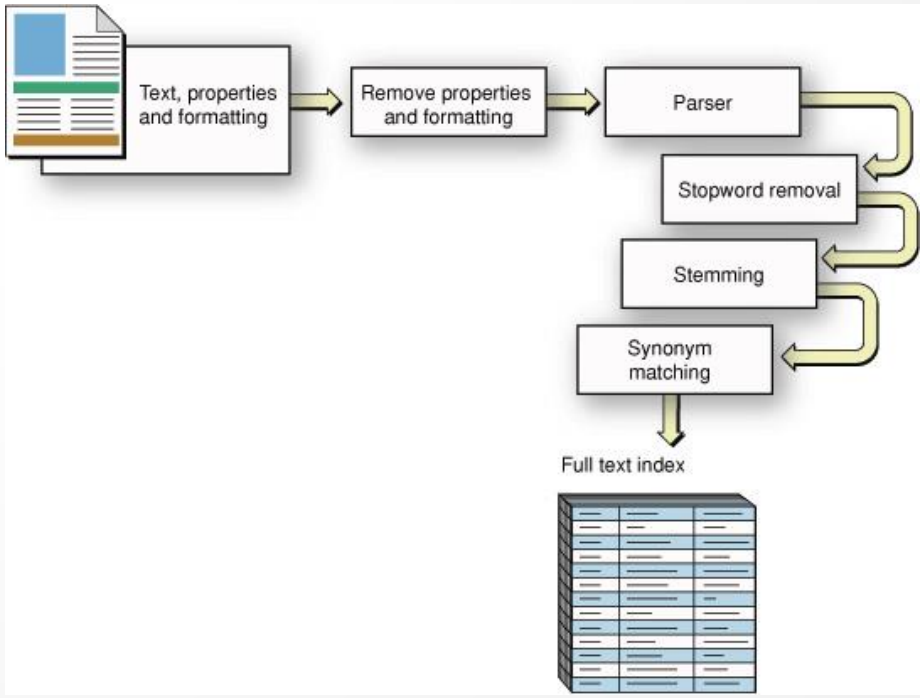
Crawler - fetching

- Buffer (frontier)
  - „to-do“ list se seznamem ještě nenavštívených (ale známých) odkazů
  - omezení počtu zpracovaných URL
- Historie
  - seznam URL, které již byly zpracovány
  - řešení proti zacyklení
  - využití při dalším zpracování zdroje
- Úložiště
  - obsahuje načtené dokumenty/stránky pro další fáze – parsování, indexování, vyhledávání

Crawler - parsing

Zdroj:
https://developer.apple.com/library/mac/#documentation/userexperience/Conceptual/SearchKitConcepts/searchKit_basics/searchKit_basics.html

- zpracování obsahu načtené stránky (dokumentu)
  - hledání dalších URL v dokumentu
  - lexikální analýza - identifikace objektů (slov) k indexování
  - stoplisting – eliminace neefektivních slov z textu (předložky, členy, apod.)
  - stemming/lematizace – standardizace slov do základního tvaru (množná čísla, zdrobněliny, předpony, apod.)
  - thezaurus – standardizace slov podle synonym ze slovníku
  - kanonizace URL – zajištění jednotnosti všech URL (velikost písmen, port, absolutní URL, PHPSESID, atd.)
- Důležitým prvkem je algoritmus řazení a ohodnocování nalezených dokumentů jejich vnitřních URL (path-ascending, focused, atd.) – určování, které URL a jak dále prohledávat



Crawler - parsing

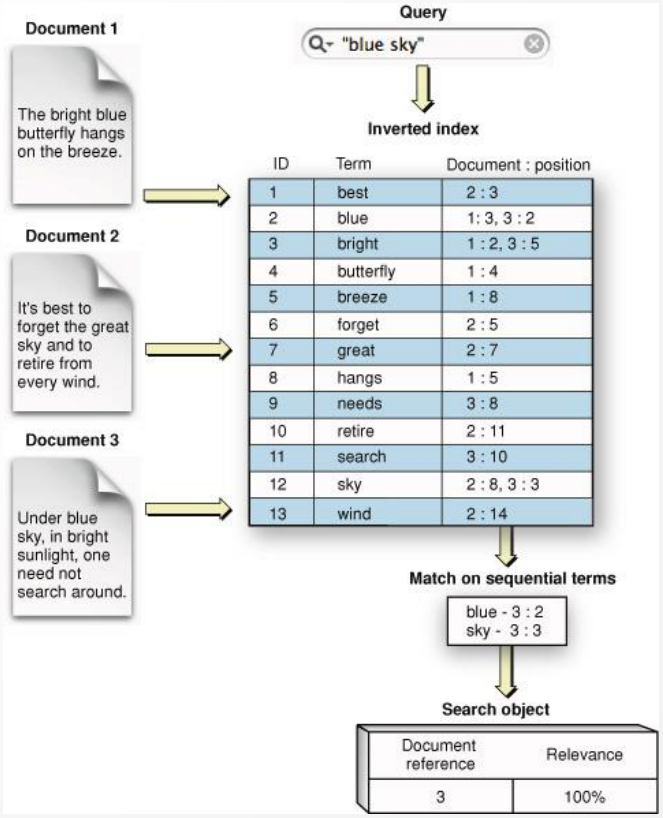
```
<p><strong>Milosevic's</strong> comments, carried by the official news agency
<em>Tanjug</em>, cast doubt over the governments at the talks, which the
international community has called to try to prevent an all-out war in the
Serbian province. "President Milosevic said it was well known that Serbia and
<a href="map.html">Yugoslavia</a> were firmly committed to resolving problems in
Kosovo, which is an integral part of Serbia, peacefully in Serbia with the
participation of the representatives of all ethnic communities," Tanjug said.
Milosevic was speaking during a meeting <br />

<img src="gb.gif" />

with British Foreign Secretary Robin Cook, who delivered an ultimatum to attend
negotiations in a week's time on an autonomy proposal for Kosovo with ethnic
Albanian leaders from the province. Cook earlier told a conference that
Milosevic had agreed to study the proposal.</p>
```

```
Milosevic comm carri offic new agen Tanjug cast doubt govern talk interna
commun call try prevent all-out war Serb province President Milosevic said well
known Serbia Yugoslavia firm commit resolv problem Kosovo integr part Serbia
peace Serbia particip representa ethnic commun Tanjug said Milosevic speak meeti
British Foreign Secretary Robin Cook deliver ultimat attend negoti week time
autonomy propos Kosovo ethnic Alban lead province Cook earl told conference
Milosevic agree study propos.
```

Indexování

Zdroj:
https://developer.apple.com/library/mac/#documentation/userexperience/Conceptual/SearchKitConcepts/searchKit_basics/searchKit_basics.html

- Data zpracována crawlerem se ukládají do databáze a vytváří se relace s URL
- Využívá se invertovaný index – setříděný seznam termů, kdy ke každému je evidována množina dokumentů
- Zároveň probíhá výpočet váhy (ohodnocení relevance a důležitosti
- mezi slovem a stránkou, SEO)
  - TF a IDF (term frequency, inverse document frequency)
  - on-page faktory (umístění slova, vzdálenost slov, klíčová slova, popisky, apod.)
  - off-page faktory (adresa stránek, zpětné odkazy, PageRank)



Vyhledávání

- Zpracování dotazu
  - tokenizace
  - parsování
  - stoplisting, stemming
  - vytvoření dotazu
  - rozšíření dotazu – thesaurus
  - ocenění termů v dotazu
  - realizace dotazu nad invertovaným indexem
  - vyhledávání nad odpovídajícími dokumenty
  - setřídění podle ohodnocení dokumentů
- Jazyková specifika
  - diakritika - transformace do unicode
  - tvarosloví
  - lokalizace stránky – heuristická analýza charakteristických slov

Hodnocení nalezených informací

- Z pohledu vyhledávače
  - frekvence výskytu termů
  - pozice termů
  - analýza vazeb (zpětné odkazy, PageRank, atd.)
  - popularita
  - datum publikování
  - velikost dokumentu vzhledem k výskytu termů
  - vzdálenost termů v dokumentu
  - význam termů vzhledem k obsahu a tématu dokumentu
  - Návštěvnost stránek a jejich popularita
  - Penalizační faktory
- Z pohledu uživatele
  - účel dokumentu a jeho typ (např. reklama vs. odborný text)
  - objektivnost, úplnost, důvěryhodnost, přesnost
  - autorství a umístění dokumentu
  - jazyková a stylistická kvalita
  - citované zdroje a reference
  - aktuálnost obsahu

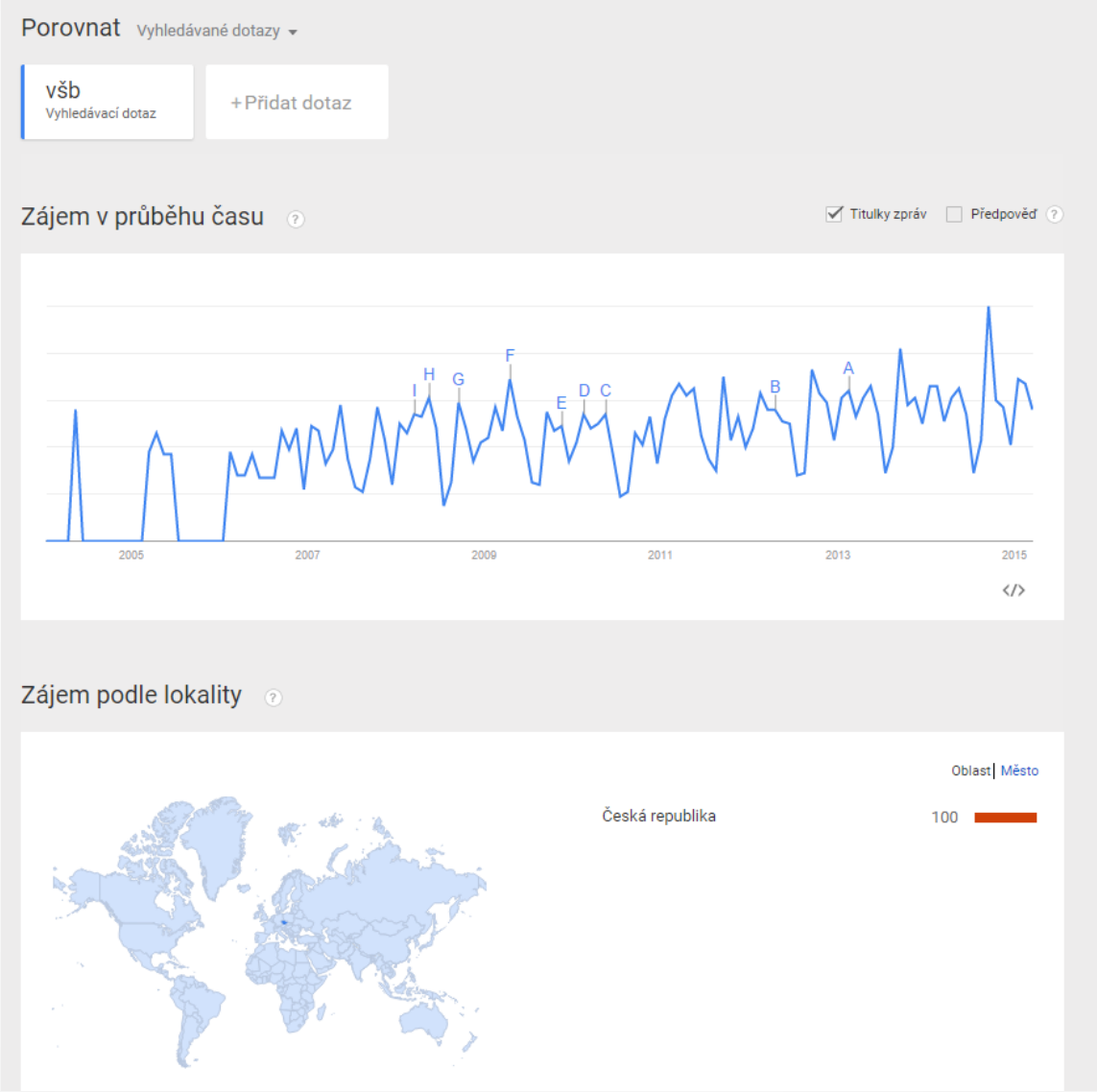
Google – pokročilé dotazy

Zdroj: http://www.inforum.cz/inforum2004/pdf/Peceny_Ondrej.pdf

http://www.google.com/patents

- josef +I.
- “zákon o účetnictví“
- brouk –volkswagen –vw
- ~help “excel”
- czechoslovakia 1950..1960
- define:orange
- notebooky filetype:xls
- intitle:medicentrum interna
- allintitle:letovy rad
- inurl:shop televize
- allinanchor:digitalni knihovna
- školení site:stk.cz
- link:www.vsb.cz
- related:www.vlada.cz
- info:www.fei.vsb.cz
- cache:www.mlp.cz spořilov
- (15/5)*2
- 10 USD in CZK
- inurl:hesla filetype:txt
- inurl:wcx_ftp.ini
- visa 4356000000000000..4356999999999999
- intitle:index.of server.at site:vsb.cz

- https://www.google.cz/trends/



Problémy dnešního vyhledávání

- Velikost indexu
  - co vše je jednotlivými vyhledávači indexováno (pokrytí)
  - vazba mezi růstem webu a indexy
- Aktualizace indexu
  - zpoždění mezi publikací informace a jejím zaindexováním
- Formáty dokumentů
  - významná část zdroje informací na internetu, která vyžaduje jiné postupy než klasické WWW stránky
- Dynamicky generované, dynamické stránky a RIA
  - stránky vzniklé na základě požadavku, které navíc mohou mít pouze dočasnou platnost
  - dynamické prvky stránek je problematické indexovat
  - technologie podpory RIA přístupů
- Index spamming
  - metody pro oklamání algoritmů pro hodnocení relevance stránek (seznamy pojmů a slov, neviditelný text, odkazy a křížové odkazy, stránky s výsledky hledání)
  - etický problém, nikoliv technologický
  - objevují se obrany ve formě penalizací „neetických" stránek
  - Google bomba - http://cs.wikipedia.org/wiki/Google_bomba

Koncepční problémy a nedostatky

- Zpracování přirozeného jazyka
  - pochopení významu slova vzhledem ke konceptu
- Interakce uživatelů a vyhledávače
  - „lidé často nemají představu o tom, co hledají“
  - správná formulace dotazu je základ úspěchu
- Ověřování informací
  - schopnost z nalezených výsledků vybrat a použít ty „správné“

Budoucnost vyhledávání

- Technické a technologické zázemí
  - zajištění platformy a algoritmů pro efektivní provoz crawlerů, indexovacích a vyhledávacích serverů, a to s ohledem na rostoucí množství informací a nové podoby jejich prezentace
- Inteligentní zpracování a tvorba dotazů
  - podpora tvorby dotazů a jejich interpretace
- Selekce zdrojů pro vyhledávání
  - vnímání zdrojů podle důvěryhodnosti
  - různé typy informací představují různé doménové oblasti
- Perzonalizace
- Integrace vyhledávačů
  - vyhledávání v rámci počítače i internetu
  - rozhraní vyhledávače je součástí aplikací
- Sémantika
  - z pohledu obsahu, indexování a hodnocení
  - z pohledu tvorby a provádění dotazů
- Sociální sítě


SEARCH
SEM
SEO
ENGINES
META
ENGINE
PAGE
RESULTS
KEYWORDS
ADVERTISING
MARKET
CONTENT
PAGES
USE
WEB
COMPANY
ATTRIBUTE
ALSO
ELEMENTS
MARKETING
EDIT
OPTIMIZATION
NAME
ONE