VIKMA06

Vyhledávání informací

14. 12. 2018: Přednáška P12+K5: Vyhledávání netextových informací, vyhledávání a internet

FF MU, podzim 2018

Mgr. Josef Schwarz
126172@mail.muni.cz

Výsledky 7. dílčího úkolu

- 7. úkol (Ptejte se knihovny)
  - Přehled dotazů
  - Statistická analýza
  - →(oba soubory v IS)

Výsledky 9. dílčího úkolu

- **Zadání:** Je dán úryvek (text z roku 1966): „*A je příznačné, že hledajíc živný kontakt s vlastní minulostí, najdou ho Francouzi ne u Crébillonů, Restifů, Laclosů a Nerciatů (příliš zdravých pro naši dobu), ale u Sada. Paulhan, Bataille, Blanchot, Beauvoirová ho opráší svými studiemi. Pak už půjde jen o to, abychom ho překonali. (Jak se to dařilo a daří, ukázal Václav Černý ve své nedávné studii.).*“ O jakou studii V. Černého jde? Zjistěte bibliografické údaje.
- **Řešilo 9 studentů, ke správnému (autorskému) řešení dospěli 2 studenti.**
- **Autorské řešení viz samostatný soubor v osnově předmětu.**

Netextové informace

- obraz, zvuk, kombinace
  - textová složka je marginální
- internet
  - velký objem netextových informací
  - omezené možnosti vyhledávání
    - vyhledávače (podle popisku – příklad 1, 2, 3)
- způsoby přístupu
  - prohlížení
  - vyhledávání

Indexace netextových inf.

- podstatně složitější než indexace textových inf.
- hlediska indexace/vyhledávání
  - hlediska 1
    - věcnost (ofness) → „tvrdá“ indexace
    - výrazovost (aboutness) → „měkká“ indexace
  - hlediska 2
    - základní vlastnosti (barva, tvar)
    - logické vlastnosti (vztah mezi objekty)
    - abstraktní vlastnosti (metaforický význam)

Vyhledávání netextových inf.

- content-based image retrieval (CBIR)
  - vyhledávání podle obsahu
  - automatické zpracování obrazu (*image processing*)
- description-based image retrieval
  - (context-based, concept-based)
  - vyhledávání podle popisu (kontextu, pojmového vyjádření) (*image indexing*)

CBIR

- vyhledávání na úrovni pixelů
  - QBIC - Query by Image Content
- objektové vyhledávání
  - extrahování obrazových objektů
- image mining (dolování obrazových informací)
  - extrakce podobných znaků z celé db
  - extrakce všech vlastností bez prvotní znalosti

Vyhledávání podle popisu

- výhoda: sémantický obsah obrazu
- nevýhoda: subjektivita→inkonzistence indexace
- způsob indexace závisí na typu kolekce a požadavcích uživatelů
- indexace
  - biografických vlastností
  - předmětových vlastností
  - fyzických vlastností
  - vztahových vlastností

Řízené slovníky pro popis netextových dokumentů

- ICONCLASS
- ATT (Art & Architecture Thesaurus)
- Thesaurus for Graphic Materials

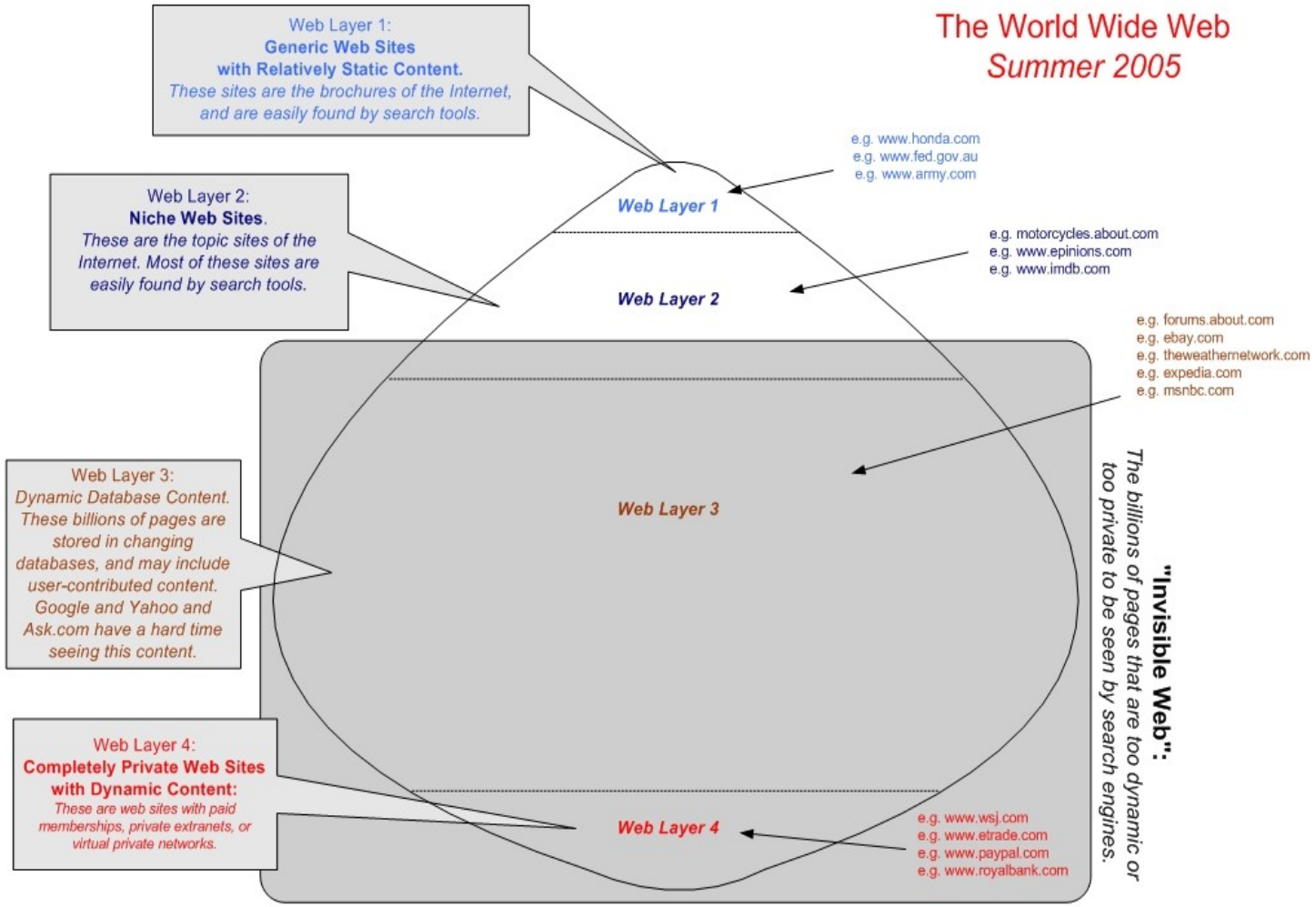
Aplikační oblasti

- průmyslové vlastnictví (ochranné známky)
- lékařství
- umění a architektura
- astronomie
- kriminologie
- …atd.

Vyhledávání a internet

- Skrytý web
- Sémantický web

Neviditelný web



Typy „neviditelnosti“

- Nepřehledný web (Opaque web)
- Soukromý web (Private web)
- Vlastnický web (Proprietary web)
- Skutečně neviditelný web (Truly invisible web)

Nepřehledný web

Obsahuje soubory, které mohou být, ale z určitých příčin nejsou vyhledávači indexované.

Důvody:

- hloubka indexování (depth of crawling)
- frekvence indexování (zprávy, inzerce, ceny akcií)
- maximální počet viditelných výsledků
- odpojené stránky

Soukromý web

Obsahuje stránky, které by robot dokázal zaindexovat, ale správce webu to znemožňuje.

- stránky chráněné heslem
- soubor robots.txt
- metatagy „noindex“, „nofollow“

Vlastnický web

Část webu, ke které je přístup pouze po splnění určitých podmínek.

- stránky vyžadující souhlas s podmínkami pro vstup
- stránky dostupné po zaplacení poplatku

Skutečně neviditelný web

Stránky, které roboty neindexují kvůli svým technickým omezením.

- dynamicky generované stránky
- relační databáze (Oracle, MS SQL Server, IBM DB2)

Přednosti hlubokého webu

- specializovaný obsah – komplexnější informace
- sofistikovanější uživatelské rozhraní
- větší důvěryhodnost
- oborovost

Přístup k hlubokému webu

- metavyhledávače
- specializované vyhledávače, katalogy, adresáře
- oborové (předmětové) vyhledávače, katalogy, adresáře
- referenční zdroje
- weby knihoven
- digitální a virtuální knihovny
- oborové databáze
- weby organizací
- knihy (archivy, e-books)
- blogy

Sémantický web

klasický x sémantický web

- Tvořen tak, aby jeho obsahu porozuměl pouze člověk
- Citlivý na použitou terminologii
- Nalezených dokumentů je obvykle příliš mnoho nebo naopak příliš málo (případně žádné)
- Výsledkem vyhledávání je pouze jedna stránka

- Rozšíření klasického webu
- Obsah ve strojově přístupné formě
- Vyhledávání podle klíčových slov nahrazeno zodpovídáním dotazů
- Dotaz je možno zodpovědět na základě extrakce informací z více stránek

Klasická podoba webu

<h1>Agilitas Physiotherapy Centre</h1>
Welcome to the home page of the Agilitas Physiotherapy Centre.
Do you feel pain? Have you had an injury? Let our staff
Lisa Davenport, Kelly Townsend (our lovely secretary)
and Steve Matthews take care of your body and soul.
<h2>Consultation hours</h2>
Mon 11am - 7pm
Tue 11am - 7pm
Wed 3pm - 7pm
Thu 11am - 7pm
Fri 11am - 3pm<p>
But note that we do not offer consultation
during the weeks of the
State Of Origin games.

Web s explicitními metadaty

- *XML + XML schéma*
- *RDF + RDF schéma*

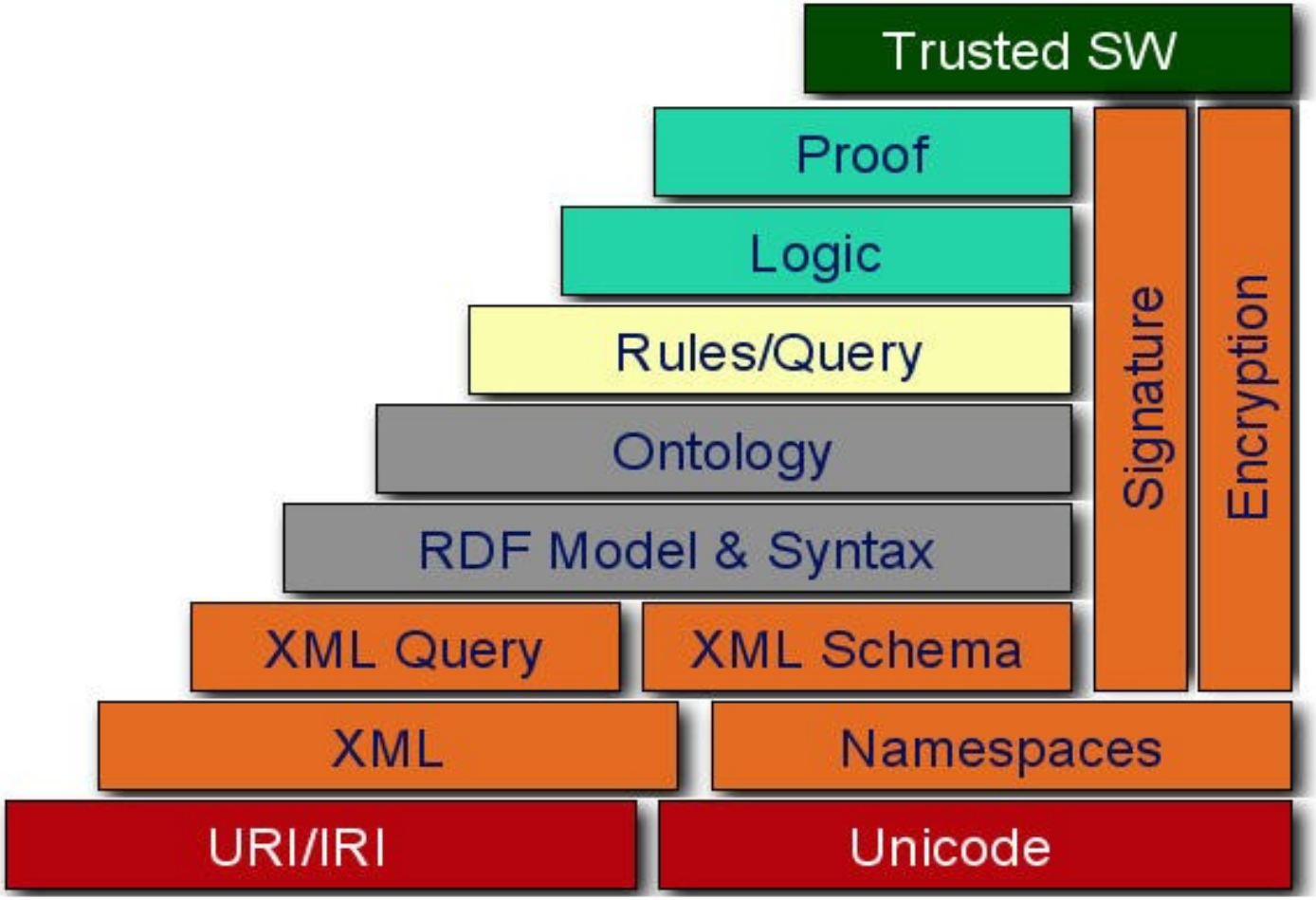
```
<company>
<treatmentOffered>Physiotherapy</treatmentOffered>
<companyName>Agilitas Physiotherapy
   Centre</companyName>
<staff>
<therapist>Lisa Davenport</therapist>
<therapist>Steve Matthews</therapist>
<secretary>Kelly Townsend</secretary>
</staff>
</company>
```

Sémantický web

- Základní složky (předpoklady) SW
  - strukturace dokumentů
  - vyjádření sémantiky - ontologie
  - vyhledávací nástroje - agenti

  - standardy
    - syntaktická složka
      - URI
    - strukturální složka
      - XML
    - sémantická složka
      - RDF + RDFS (schéma RDF)
      - OWL, OIL

Vrstvy sémantického webu



Sémantický web – příklady řešení

- W3C
- příklad aplikace RDF
  - energetika
- Výzkum
  - The Open University London, Knowledge Media Institut
    - Magpie
  - Stanford Knowledge Systems Laboratory
    - DAML (agenti)
  - EU, 5. rámcový program
    - On-to-knowledge