Vyhledávání multimediálního obsahu na Internetu

Michal Krsek, UISK UK & CESNET
Ivan Doležal, CESNET
Michal Illich, Jyxo

Elektronická média

- TV & rádio
- Organizována v kanálech/programech
- Nulová demokracie při šíření (programování managementem stanic)
- Centralizovaná produkce

Internet

- Není pouze Web (audio/video a další obsah)
- IPTV / Video on demand
- Navigace především prostřednictvím webu

=> není (prakticky) možné vyhledávat jinými prostředky/protokoly

Možnosti vyhledávání I

- Rozpoznávání hlasu
  - Rozpoznání jazyka
  - Akcenty
  - Nemluvený zvuk
- Rozpoznávání videa
  - Interpretace dotazu (bush vs. Bush)
  - Nízká kvalita videa

Možnosti vyhledávání II

- Indexování webových stránek (odkazů)
  - Dělá Yahoo! (cíl pro google bombu)

Metadata

  - “Vnější Metadata” (knihovnické paradigma)
  - Metadata v souborech (vložena během postprodukce)

Popis projektu

- Ustaven v roce 2003
- “Google pro audio a video na Internetu”
- Bez podpory vlastníků obsahu
- Modulární návrh
- Počátek s indexací .cz

Technický popis I

- Crawler
  - Prochází web a získává adresy (URL)
  - Předává adresy multimediálních souborů
  - Software vyvíjený Jyxo (Linux aplikace)

Technický popis II

- Distiller
  - Importuje adresy multimediálních souborů
  - Získává metadata (vytváří XML soubory)
  - Vytváří náhledy (pokud je video v souboru)
  - C# software a mplayer (windows aplikace)

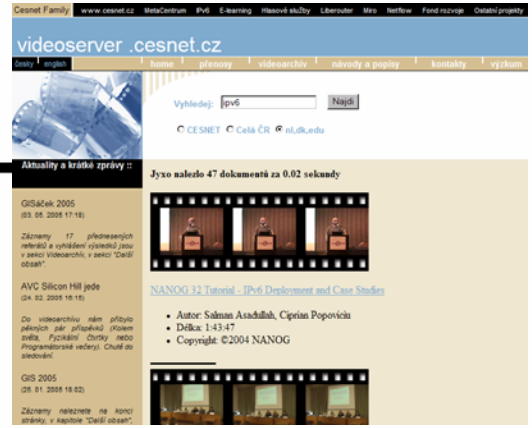
Technický popis III

- Databáze
  - Jako zdroj používá XML soubory
  - Odpovídá na back-end dotazy z webů
  - Jazykové přizpůsobení (lemmatizace, stop slova)
  - Software vyvíjený Jyxo

crawling

Prochází webový prostor

Získává adresy
Dál předává jen adresy A/V souborů

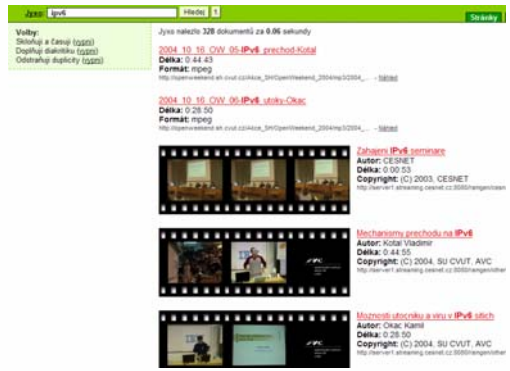


destilace

Získává metadata
z A/V souborů

indexing
hledání

www.
yournamehere.
edu



Vytváří a obsluhuje fulltextovou databázi
Odpovídá na back-end dotazy webů

Aktuální stav I

- HW
  - 10 crawlerů (<40.000 Kč)
  - Prostor na úložném serveru
  - Server pro fulltext DB (<40.000 Kč)
  - Stanice pro destilaci (X kancelářských PC)
  - Připojeno 1 Gb/s do CESNET2

Aktuální stav II

- Database
  - .cz, .sk, .hu, .de, .fr, .nl, .it, .dk, .edu, .pl, .pt, .ch, .ua
  - > 3.100.000 adres
  - > 2.100.000 validních
  - ~ 600.000 s náhledy
  - Náhledy ~2.500.000 (=17 GB na disku)

Ukázka?

Další vývoj

- Evoluční
  - Další domény
  - Podpora dalších jazyků
- Revoluční
  - Implementace knihovnického přístupu (spolupráce vydavatelů obsahu) pomocí metod OAI
  - Detekce duplicit

Adresy systému

- URL
  - http://multimedia.jyxo.cz
  - http://videoserver.cesnet.cz/videoarchiv_en.php

- Zájemci o přístup k XML rozhraní žádejte e-mailem

Otázky ?
Komentáře ?

Michal Krsek, Michal.Krsek@cesnet.cz (akademická obec, výzkumná spolupráce)
Michal Illich, michal@illich.cz (obchodní sféra)