Vyhledávání multimediálního obsahu na Internetu

Michal Krsek
CESNET, z.s.p.o. a UISK Karlova Univerzita, Praha
Michal.Krsek@cesnet.cz

Ivan Doležal
CESNET, z.s.p.o., Praha
Ivan.Dolezal@vsb.cz

Michal Illich
Jyxo, s.r.o., Praha
illich@jyxo.cz

INFORUM 2006: 12. konference o profesionálních informačních zdrojích
Praha, 23. – 25. 5. 2006

***Abstrakt**: Objem multimediálního obsahu na Internetu stále roste. Pro vyhledávání v multimediálních souborech není možné používat plnotextové vyhledávání, proto je třeba pracovat s metadaty. Sdružení CESNET a společnost Jyxo vyvíjí vyhledávač multimediálních dat na Internetu, který pracuje s metadaty, které získává z Internetu. Článek představí principy, kterými se vyhledávač řídí.*

Motivace

S rozvojem širokopásmového přístupu k Internetu se zvětšují možnosti uživatelů využívat i pokročilejší formy multimediálního obsahu, například audio a video. Plnohodnotné využití potenciálu těchto služeb na Internetu uživatelem vyžaduje možnost vyhledávání obsahu. Požadavky uživatelů směřují k potřebě vyhledávat ve zdrojově heterogenních bázích dat při použití jednoduchého rozhraní.

Hledání cest k efektivnímu a zároveň uživatelsky jednoduchému vyhledávání multimediálních dat je cílem našeho projektu.

Současný stav

S rozvojem širokopásmového připojení si začaly důležitost Internetu uvědomovat velcí majitelé obsahu. Jejich portály jsou ovšem postaveny pro uživatele pasivně konzumujícího televizní program, který se pohybuje pouze v rámci jednoho poskytovatele. Pokud obsahují vyhledávání, pak pouze v rámci jednoho portálu.

Protipólem k velkým mediálním koncernům existuje množství uživatelů, kteří svůj audio a video materiál vystavují jako obohacení svých stránek. Vyhledávání v těchto materiálech je klasickými metodami prakticky nemožné.

Podobná situace existuje ve světě WWW (respektive HTML), nicméně pro WWW existují vyhledávací stroje, které umožňují vyhledávání dat založených na textové informaci. Tradiční vyhledávače spoléhají na provázanost informací na webových stránkách s obsahem multimediálních souborů. Služba Google video vyhledává pouze v souborech, které jsou uloženy přímo na serverech Google a služba Yahoo! Video vyhledává pouze v URL.

Provozovatelé mediaték či obchodů s multimediálním obsahem disponují vyhledávacími nástroji. Tyto nástroje jsou omezeny na konkrétní archív a jejich použitelnost v heterogenním prostředí je sporná. Aktivity veřejných knihoven v oblasti digitálního zpracování a výměny informací jsou soustředěny na práci s tištěnými materiály.

Problémem, které řešíme, se zabývá několik start-upů, příkladem může být společnost Truveo, která je od přelomu roku součástí společnosti AOL.

Návrh řešení

Vyhledávání v audio a video souborech je možné několika způsoby.

Prvním způsobem je porovnávání obsahu s vyhledávaným vzorem (například slovo vůči zvukovému záznamu, konkrétní obrázek vůči filmu nebo text vůči titulkům). Tento způsob vyhledávání vzhledem k Internetu nelze v současné době díky nízké kvalitě záznamů, nízkému relativnímu výkonu vyhledávacích algoritmů na běžně dostupných zařízeních a heterogenitě materiálu (velké množství kodeků a formátů) aplikovat. Specifickým problémem je uživatelské rozhraní – uživatelé pokládají dotazy v textové formě, kterou je v případě porovnávání obsahu třeba interpretovat. V případě jednoduchých výrazů jde o vytvoření rozsáhlé databáze vzorů (obrázky politiků), v případě abstraktních slov (politika, IPv6) nelze takový vzor vytvořit. Fakticky je nutno vytvořit tezaurus termínů vůči obrázkům.

Druhým způsobem je vyhledávání v metadatech, což jsou textová data, která jsou uložena tak, aby byla dostupná současně s vlastním materiálem. V prostředí Internetu převažují metadata uložená přímo v multimediálních souborech, respektive na webových stránkách, které na příslušné soubory ukazují. Textovou informaci je potom možné zpracovat algoritmy plnotextového vyhledávání. Pro plnotextové vyhledávání je dnes k dispozici velké množství software (včetně balíků dostupných zdarma). Zvolili jsme formu spolupráce s plnotextovým vyhledávačem Jyxo (řešitelský tým nemusel řešit běh vyhledávače a front-end pro uživatele). Spolupráce s běžícím systémem nám také umožnila získat dostatečně široký objem materiálu k vyhledávání.

Popis systému

Systém je tvořen standardními komponentami plnotextového internetového vyhledávače (crawler, indexer, databáze/vyhledávač), se kterými je integrována komponenta destilator, která získává metadata z definovaných multimediálních souborů. Tato komponenta komunikuje off-line s ostatními komponentami systému standardními protokoly (ssh/scp) rozhraními (čistý text a XML) je snadno integrovatelná do jakéhokoliv prostředí.

Komponenta crawler ze stránek, které získá procházením webového prostoru, uloží URL audio a video souborů (filtr je nastaven na přípony souborů a typy médií poskytované serverem) do textového souboru. Tento soubor je následně importován do SQL databáze. Destilator prochází adresy v databázi a ze získaných metadat (proces destilace) vytváří XML soubory, které umisťuje do výstupního adresáře. Z tohoto adresáře jsou protokolem SCP přenesena do systému, kde běží indexer, který z dat vytváří běžnou plnotextovou databázi, nad kterou uživatelé vyhledávají.

Jak procházení webu, tak následná destilace jsou časově náročný proces, který není možné provozovat v jedné instanci. Komponenty systému jsou provozovány v paralelním režimu a jejich komunikace je asynchronní.

Plnotextová databáze poskytuje XML výstup, jenž používají webové portály, prostřednictvím kterých kladou systému dotazy uživatelé.

Destilator

Klíčovou komponentou systému je destilator. Vzhledem k potřebě indexovat co nejširší spektrum formátů a kodeků (a dynamickému vývoji v této oblasti) jsme upustili od vývoje vlastního dekodéru. V průběhu vývoje jsme vyzkoušeli několik jednoúčelových utilit dostupných volně na Internetu, nicméně se nám nepodařilo získat uspokojivou kvalitu dat a stabilitu systému.

Výsledná podoba destilatoru je Win32 aplikace psaná v jazyce C#, která předává jednotlivé adresy ActiveX objektům, které jsou součástí multimediálních přehrávačů (Real One Player, Windows Media Player, QuickTime player a další). Tyto objekty se posléze pokoušejí otevřít adresy některým z kodeků nabízených operačním systémem (WM) nebo dodávaných pro přehrávač RealOne Player.

Snímání obrázků je realizováno programem mplayer, který dokáže uložit snímek obrazovky do souboru. Vzhledem k tomu, že snímky jsou uloženy v originální velikosti, je potřeba snímky transformovat do vhodné velikosti a formátu. To se děje dávkově při předávání dat mezi destilatorem a plnotextovou databází.

Směry dalšího rozvoje systému

V průběhu řešení a provozem pro vědeckou, výzkumnou a akademickou komunitu jsme objevili možné směry dalšího vývoje.

Prvnímu směru je vyvolán faktem, že vlastníci souborů často vložená metadata nevyplňují. Spoléhají pravděpodobně na to, že materiál bude dostupný pouze z jejich WWW portálu, případně jde z jejich strany o opomenutí při publikaci příspěvků. Tento přístup není v silách řešitelů změnit, nicméně řešení spočívá v doplnění funkcionality systému. Předpokládáme kombinaci současného přístupu bez potřeby kooperace s vlastníky obsahu se získávání metadat od provozovatelů obsahu pomocí metod OAI (Open Archives Initiative).

Druhý směr vývoje určuje potřeba identifikovat obsahové duplicity. Stejný obsah může být (a je) publikován jeho vlastníky v různých kvalitách, formátech, rozlišeních a na různých serverech, přičemž jednoduché porovnání obsahu například pomocí kontrolních součtů není možné. Předpokládáme, že vyvineme systém hodnotící podobnost souborů na základě informací, které o sobě soubor nese. Současně předpokládáme vytvořit systém identifikace materiálů (buď adopcí některého z existujících handle systémů nebo vytvořením vlastního).

Prozatím posledním směrem vývoje je problém značné rozptýlenost multimediálního obsahu do malého množství (méně než procento) serverů. Zatímco vyhledávače webu potřebují skutečně projít a zaindexovat všechny stránky, v případě multimediálních souborů tato potřeba není. Na webových stránkách jsou z hlediska projektu zajímavé pouze adresy multimediálních soborů – a počet použitelných odkazů je nižší než jedno promile (konkrétní číslo je různé pro různé země) adres stránek, které crawler prochází. Proto je naším cílem vyvinout inteligentní crawler, který bude schopen procházet pouze takové servery, u který je velká naděje na získání použitelných adres.

Zhodnocení projektu

Výsledkem projektu je funkční fulltextový vyhledávač v multimediálních datech dostupných na českém Internetu, běžící na adrese http://multimedia.jyxo.cz/, což je vzhledem k plánovaným výsledkům plné naplnění cílů. Systém používáme i k vyhledávání ve videoarchívu CESNETu (http://videoserver.cesnet.cz) a nabízíme ho k volnému použití všem akademickým a výzkumným organizacím.

V dubnu 2006 indexujeme Internet ČR, Dánska, Francie, Itálie, Maďarska, Nizozemí, Polska, Portugalska, Slovenska, SRN, Švýcarska, Ukrajiny a akademickou doménu .edu. Indexování dalších národních domén je závislé na zájmu akademických komunit z dotyčných zemí. Indexování rozsáhlých generických domén (například .com) je závislé na vývoji komponenty inteligentní crawler.

Celkem pracujeme s více než 2.500.000 adresami multimediálních zdrojů. Byli jsme schopni vydestilovat metadata z 2.000.000 adres (zbytek tvoří nevalidní adresy nebo jde o chráněný obsah). Náhledy jsme získali u více než 350.000 adres (zbytek tvoří audiosoubory nebo soubory zakódované kodeky, které destilator nezná). Z každého souboru se snažíme získat tři náhledy, které ukládáme ve dvou formátech (PNG a JPG v rozlišení 96x72 pixelů). Celkem máme uloženo více než 2.100.000 souborů s náhledy, jejichž souhrnná velikost je 15 GB.

Příloha

Formát souboru pro předávání dat mezi destilátorem a plnotextovou databází

```
<!--
    File:  destilator-0-3.dtd
```

```
 Purpose:   Metadata destilator interchange format
 Version:   0.3 2003-12-01
Location:  http://prenosy.cesnet.cz/dtd/

Basic structure:

<assets>
<file
    URL="url" - URL to the file
    streamable="(0|1)" - indicates if media file is streamable
    reachable="(0|1)" - is this asset accessible
    format="text" - media file format
/>
<title>Title of the asset (extracted from metadata)</title>
<authors>Authors of the asset (extracted from metadata)</authors>
<copyright>copyright holders of the asset (extracted from
metadata)</copyright>
<length>length of the asset - 1:00:00 / 0 (for infinite)</length>
<islive>indicates if the media is live (0|1)</islive>
<description>description of the asset (extracted from
metadata)</description>
<keywords>keywords in the asset (extracted from metadata)</keywords>
<rating>rating of the asset (extracted from metadata)</rating>

<stream>
<codec>codec identification (plain text)<codec>
<bitrate>bitrate</bitrate>
<media>identifies stream payload - audio/video/pictures ... others</media>
<sampling>sampling rate (only for sound)</sampling>
<width>width of screen (only for picture/video)<width>
<height>height of screen (only for picture/video)</height>
<fps>frames per second (only for video)</fps> - pocet snimku za vterinu
pouze pro obraz
</stream>

more <stream> .... </stream> records

</file>

more <file> .... </file> records

</assets>

-->

<!ENTITY % zeroone
      "(0|1)"
>

<!-- top level labels -->
<!ELEMENT assets (file*)>
<!ELEMENT file (title?, authors?, copyright?, length?, islive?,
description?, keywords?, rating?, stream*)>
<!ATTLIST file
          URL CDATA #REQUIRED
          streamable %zeroone; #REQUIRED
          reachable %zeroone; #REQUIRED
          format CDATA #REQUIRED
>
<!ELEMENT title (#PCDATA)>
<!ELEMENT authors (#PCDATA)>
```

```
<!ELEMENT copyright (#PCDATA)>
<!ELEMENT length (#PCDATA)>
<!ELEMENT islive (#PCDATA)>
<!ELEMENT description (#PCDATA)>
<!ELEMENT keywords (#PCDATA)>
<!ELEMENT rating (#PCDATA)>
<!ELEMENT stream (codec, bitrate, media?, sampling?, width?, height?,
fps?)>
<!ELEMENT codec (#PCDATA)>
<!ELEMENT bitrate (#PCDATA)>
<!ELEMENT media (#PCDATA)>
<!ELEMENT sampling (#PCDATA)>
<!ELEMENT width (#PCDATA)>
<!ELEMENT height (#PCDATA)>
<!ELEMENT fps (#PCDATA)>


<!--End of (destilator-0-3) Definition-->
```