# VYHLEDÁVÁNÍ V MULTIMEDIÁLNÍCH DATECH HETEROGENNÍCH SÍTÍCH A NA INTERNETU

*Ivan Doležal – Michal Krsek, CESNET ,Michal Illich, Jyxo*

**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. Překlenutí této bariéry je cílem našeho projektu.

Počet veřejně dostupných audio a video souborů, které jsou pod veřejným URL uloženy v naší databázi, dosahuje aktuálně (srpen 2005) objemu 1 000 000 příspěvků.

**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. Pokud je nám známo, podobný systém pro vyhledávání audio a video materiálu není v běžném provozu. Služba Google videosearch vyhledává ve skrytých titulcích a služba Yahoo! Video vyhledává pouze v URL.

Na Internetu existují rozsáhlé peer-to-peer sítě tvořené aplikacemi určené ke sdílení uživatelů mezi sebou. Tyto sítě obsahují vyhledávací mechanismy jako nedílnou součást své funkčnosti a proto nejsou cílem projektu.

**Návrh řešení**

Vyhledávání v audio a video souborech je možné dvěma způsoby.

Prvním způsobem je porovnávání obsahu s vyhledávaným vzorem (například slovo vůči zvukovému záznamu, 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) je nutné zvolit nejbližší konkrétní obraz. V případě víceslovního dotazu je nutné vzory kombinovat.

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 analogicky k plnotextovému vyhledávání. Pro plnotextové vyhledávání je dnes k dispozici velké množství software (včetně balíků dostupných zdarma). My jsme zvolili formu spolupráce s plnotextovým vyhledávačem Jyxo (řešitelský tým nemusel řešit běh systému 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, front-end), 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 WWW, uloží URL audio a video souborů (filtr je nastaven na přípony souborů a content-type poskytované serverem) do textového souboru (každé URL jeden řádek). Tento soubor je následně protokolem SCP přenesen na server, kde k němu má přístup destilator. Destilator při zpracování souboru prochází jednotlivá URL a z nalezených metadat vytváří XML soubory (formát viz. příloha), 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.

**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 jednotlivá URL ActiveX (OLE) objektům, které jsou součástí multimediálních přehrávačů (Real One Player, Windows Media Player, QuickTime player). Tyto objekty se posléze pokoušejí otevřít URL některým z kodeků nabízených operačním systémem (WM) nebo dodávaných pro přehrávač RealOne Player. Data získaná porovnáním výstupů z obou objektů jsou pak transformována do formátu XML.

Dostupnost materiálu a jeho korektní formát řeší přehrávač (v případě, že soubor nelze načíst, vrátí ActiveX objekt chybový stav).

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 shodné velikosti a formátu. To se děje dávkově při předávání dat mezi destilatorem a plnotextovou databází.

Vzhledem k otevřeným vstupům a výstupům může být destilator nasazen do prakticky jakéhokoliv plnotextového vyhledávače na Internetu.

Vzhledem k velkému množství URL běží destilace na vícero počítačích, jsou URL uložena v relační databázi a všechny počítače, na kterých běží destilator, pracují s touto databází.

**Problémy**

V průběhu řešení problému jsme objevili tři problémy, které částečně omezují použitelnost systému.

Prvním problémem je fakt, že vlastníci souborů často 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.

Druhým problémem je nestabilita ActiveX objektů v případě, že kodek zvolený pro přehrávání multimediálních dat narazí na takovou jejich variantu, s níž si není schopen korektně poradit. V 10% případů destilator zamrzne. Problém jsme vyřešili aplikací, která destilator ukončí v případě jeho zamrznutí.

Třetím problémem je omezené množství informací nabízené ActiveX objekty a jeho nekvalitní implementace. Přehrávače nabízejí pomocí Acti-

veX rozhraní pouze podmnožinu metadat, která jsou v multimediálních souborech uložena. Předávané informace jsou navíc zkreslující – příkladem může být informace o datovém toku předávaná RealOne Playerem. Prostřednictvím ActiveX rozhraní lze získat pouze údaj odpovídající součtu datové rychlosti všech proudů formátu SureStream, nikoliv už údaje o počtu toků a jejich jednotlivých rychlostech, navzdory faktu, že API pro získání tohoto údaje je v dokumentaci uváděno několik let.

**Zhodnocení projektu**

Výsledkem projektu je funkční fulltextový vyhledávač v multime-diálních datech dostupných na českém Internetu, běžící na adrese http://www.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.

```
<!--
    File:  destilator-0-3.dtd
 Purpose:   Metadata destilator format
 Version:   0.3 2000-12-01
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 meta-
data)</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 meta-
data)</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?, descrip-
tion?, 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-->
```