Inteligentní vyhledávání na WWW

Richard Botík[1], Luboš Popelínský[2]

[1] Právnická fakulta, Masarykova Universita Brno, botik@law.muni.cz
[2] Fakulta informatiky, Masarykova Universita Brno, popel@fi.muni.cz

Abstrakt AVS system tryies to improve AltaVista web search engine. The task of document ordering according to their relevance is solved by means of an expert system. The knowledge base contains knowledge on relations of particular tags and knowledge on their importance as well as information on relevance of key words. In 89% test cases the ordering given by AVS was at least as good as that one given by AltaVista.

1 Vyhledávání informací na WWW

Vyhledávače informací na WWW (anglicky Search Engines) [2] patří mezi systémy, které na základě předem vytvořeného indexu vyhledávají dokumenty podle požadavků uživatele. Způsoby vyhledávání informací pomocí těchto služeb lze v podstatě rozdělit do dvou kategorií – vyhledávání podle klíčových slov a vyhledávání v tématicky kategorizovaných katalozích URL. Každý z vyhledávačů umožňuje alespoň jeden způsob vyhledávání, některé dokonce oba. Mezi všemi dominuje silná pětka vyhledávačů ve složení AltaVista, Excite, Hotbot, Infoseek a Lycos. Všechny jsou typickými představiteli vyhledávání podle klíčových slov, nabízejí však i druhý způsob vyhledávání, pouze AltaVista je 'specialistou' na vyhledávání podle klíčových slov. Typickým představitelem vyhledávání v tématicky kategorizovaných katalozích URL je Yahoo.

2 AltaVista

AltaVista <http://altavista.digital.com/> se specializuje na vyhledávání podle klíčových slov. Svým rozsahem zpracovaných informací se řadí ve své třídě na přední místo, ne-li úplně na to nejvyšší. Projekt AltaVista byl započat v Palo Alto v Kalifornii v Digital's Research Laboratories v létě roku 1995. Spojením rychlého procházeče Webu s výkonným indexátorem bylo možno již na podzim 1995 vytvořit obrovský index Webu. Již po 2 měsících testování byl vytvořen do té doby zatím největší index Webu z více než 16 milionů WWW stránek. Do veřejného užívání byla AltaVista dána 15. prosince 1995. První den provozu bylo napočítáno kolem 300 000 přístupů. Tento počet se během prvních tří týdnů provozu zvýšil až na 2 000 000 a za 9 měsíců od uvedení do provozu dosahoval úchvatných 19 000 000 přístupů za den. Výkonné 64 bitové procesory Alpha spolu s výkonným softwarem jsou schopny najít odpověď na jakýkoli dotaz v době kratší než je jedna sekunda, a to i přesto, že velikost indexu je 100 GB (konec roku 1997). Kromě vyhledávacího stroje na výše uvedené adrese má nyní AltaVista několik zrcadel po celém světě. Pro Českou republiku je nejvýznamnější zrcadlo na adrese http://altavista.telia.com/, které uživatelům z ČR vychází vstříc českým prostředím.

Klady

1. Velkou výhodou je rychlost zpracování dotazů
2. Velké množství nalezených dokumentů => je velká pravděpodobnost, že námi hledaná informace byla skutečně nalezena
3. Možnost doladění dotazu
4. Vyhledávání dokumentů napsaných v určitém jazyce

Zápory

1. Průměrně 15% z nalezených dokumentů již neexistuje (viz. Tab. 1).
2. Velké množství nalezených dokumentů => příliš mnoho balastu.
3. Především v případě jednoslovných dotazů (zadáno pouze 1 klíčové slovo) je častým jevem skutečnost, že v nalezeném dokumentu se zadané slovo vyskytuje pouze jako součást hypertextového odkazu, nikoli však v samotném textu.
4. Řazení nalezených dokumentů, a to i přes značné nedávné zlepšení, je v řadě případů ne příliš kvalitní.

AVS (**A**lta**V**ista **S**ort) je asistent vyhledávače AltaVista, jenž se snaží odstranit nedostatky AltaVisty. V následujícím textu bude popsáno jeho ovládání a použitý algoritmus Na závěr budou předvedeny výsledky a diskutovány možnosti dalšího rozšíření.

3 AVS

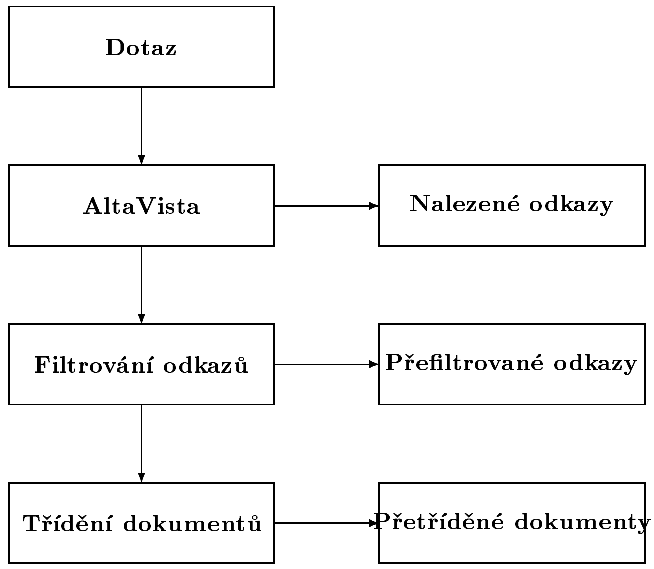
3.1 Základní schéma

Činnost systému AVS probíhá v několika krocích podle následujícího schématu (viz Obr. 1): uživatel zadá dotaz ve formě klíčových slov, který syntaxí musí odpovídat dotazu položeném v jednoduchém režimu vyhledávače AltaVista. Zadaný dotaz předá AVS vyhledávači AltaVista a získá od něj požadovaný počet odpovědí. Odpovědi zpracuje a zobrazí. Následující krok odstraní všechny neexistující a momentálně nedostupné odkazy. Přefiltrovaný seznam odkazů je opět zobrazen. V posledním nejdůležitějším kroku celého systému dojde k přetřídění nalezených dokumentů podle relevance.

3.2 Redukovaný výpis odkazů

Po získání odpovědí AVS zpracuje získané výsledky, které vypíše v následujícím tvaru:

- Z každého WWW serveru je zobrazen pouze 1 odkaz, a to ten, jenž byl AltaVistou vrácen jako nejlepší z tohoto serveru, tzn. že ze všech odkazů na tento server byl zobrazený odkaz vrácen jako první.
  Výhoda: zredukuje se množství vypisovaných odkazů
- Je-li mezi nalezenými odkazy na dokumenty více odkazů na týž WWW server, zobrazí se malá ručka spolu s nápisem *Více nalezených odkazů z toho serveru*. Kliknutím na ručku nebo na nápis budou zobrazeny všechny nalezené odkazy z tohoto serveru.



Obrázek1. Schéma systému AVS

Pro zjištění relevance dokumentů jsme zvolili expertní systém, jenž jsme implementovali s pomocí shellu pro tvorbu produkčních systémů CLIPS [3]. Kromě znalostí hypertextové struktury dokumentů jsme využili zkušenostmi nabytých znalostí o vztazích jednotlivých částí HTML dokumentů a polohy slov k relevanci dokumentů.

3.3 Odstranění neexistujících a momentálně nedostupných odkazů

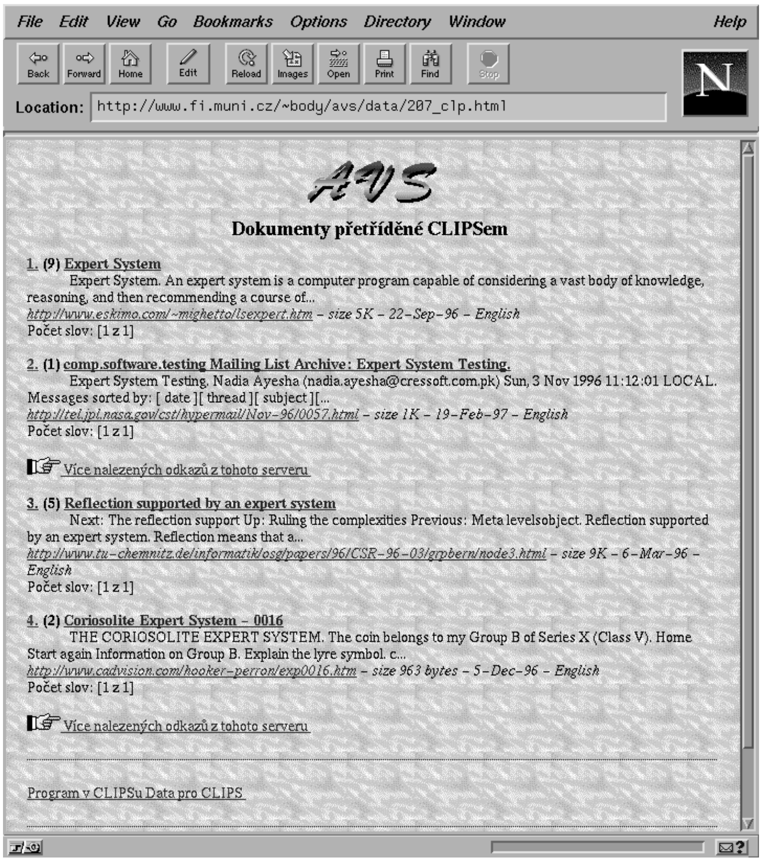
Z praxe ověřeným faktem je skutečnost, že asi 15% z nalezených odkazů je již neplatných, tzn. že neukazují na žádný dokument. Takovéto odkazy nemají žádný význam a mohou být vyřazeny. Jiným faktem je skutečnost, že se často nelze v danném časovém limitu spojit se vzdáleným počítačem či načíst požadovaný dokument. Proto také odkazy, které jsou momentálně nedostupné, a tudíž momentálně nezajímavé, mohou být též ze seznamu nalezených odkazů vyřazeny. Tento krok systému AVS tedy provede odzkoušení všech nalezených odkazů a vyřadí z nich ty, které jsou již neplatné nebo jsou momentálně nedostupné. Po dokončení kontroly zobrazí systém AVS výsledky ve stejném tvaru jak je popsáno v odstavci 3.2, obsahující pouze „živé“ odkazy. Všechny odkazy, jež bylo nutno odstranit, jsou uchovávány a lze je získat pouhým kliknutím na odkaz s nápisem **Nedostupné a neexistující odkazy**.

3.4 Po přetřídění odkazů expertním systémem

Tento krok je nejdůležitějším, ale také časově nejvíce náročným krokem systému AVS. Každá WWW stránka, jejíž URL je mezi nalezenými odkazy, je systémem AVS načtena a zanalyzována. WWW stránky, které nemohly být načteny, jsou momentálně nedosažitelné a dále s nimi není pracováno. Jakmile je načítání stránek dokončeno, provede se samotné přetřídění. Výsledkem je seznam odkazů ve stejném tvaru jako v předchozích případech, zbavený neexistujících a momentálně nedostupných odkazů a v pořadí určeném systémem AVS. Obr. 2 ukazuje příklad takového výstupu. Dokumenty jsou seřazeny v pořadí určeném systémem AVS. Za každým pořadovým číslem následuje v závorce pořadí dokumentu podle AltaVisty. Na konci údajů o dokumentu je v hranatých závorkách uveden skutečný počet různých klíčových slov obsažených v dokumentu (např. [3 z 5] značí, že tři slova z pěti zadaných jsou v dokumentu obsažena). Kliknutím na číslo udávající pořadí dokumentu lze okamžitě získat aktuálně načtenou kopii WWW stránky.

4 Algoritmus

HTML dokumenty se skládají z celé řady HTML konstrukcí (dále tagů). Každý z těchto tagů má pro dokument jiný význam, ať již vizuální nebo funkční. Proto lze jistým způsobem tyto tagy uspořádat podle významnosti pro HTML dokumenty. Ovšem zkonstruovat takovéto uspořádání je problém obecně velmi složitý, neboť nelze objektivně určit, zda tag A je důležitější než tag B.



Obr. 2 Přetřídění odkazů

Pro účely porovnávání dokumentů jsme vybrali 7 tagů, které mají podle našich zkušeností větší význam pro důležitost dokumentu, a ohodnotili je. Těmito tagy jsou *TITLE, H1, H2, H3, B, I, A* s následujícím významem:

TITLE	Titulek stránky
H1	Nadpis největší velikosti
H2	Nadpis menší než v případě H1
H3	Nadpis menší než v případě H2
B	Text uvedený uvnitř tohoto tagu bude vypsán **tučně**
I	Text uvedený uvnitř tohoto tagu bude vypsán *italikou*
A	Definice hypertextového odkazu

Aby bylo možno určit pořadí jednotlivých dokumentů, je třeba určit jejich relevanci, podle které budou posléze uspořádány.

K určení relevance a následně celkového pořadí jsme použili expertní systém AVS, do něhož jsme zanesli znalosti ve tvaru faktů a pravidel o struktuře HTML dokumentů, důležitosti jednotlivých tagů, vztahu mezi jednotlivými tagy a významnosti polohy slov v dokumentu k relevanci. ES pracuje následujícím způsobem:

1. Zjištění počáteční relevance dokumentů (blíže viz [1])
2. Uplatnění znalostí a zjištění výsledné relevance
3. Uspořádání dokumentů podle takto získané relevance

Znalosti, které se v expertním systému AVS používají, jsou založeny na vztazích tagů k dokumentu, dále na vztahu polohy klíčových slov v dokumentu k relevanci dokumentu atp. Použité znalosti jsou následující:

– Klíčové slovo vyskytující se zároveň v tazích TITLE a H1 zvyšuje relevanci dokumentu.
– Klíčové slovo vyskytující se pouze v tagu A relevanci dokumentu snižuje.
– Klíčové slovo nacházející se v první čtvrtině dokumentu zvyšuje relevanci dokumentu.
– Klíčové slovo nacházející se v poslední čtvrtině dokumentu snižuje relevanci dokumentu.
– Obsahuje-li tag pouze zadaná klíčová slova, pak se relevance dokumentu zvýší (je zkoumán pouze tag TITLE).
– Obsahuje-li dokument hypertextový odkaz na jiný dokument, pak se zvýší relevance dokumentu, na který odkaz ukazuje.

Po uplatnění všech výše uvedených znalostí dostaneme pro každý dokument jeho relevanci R vzhledem k zadanému dotazu. Pokud mají dva dokumenty stejnou relevanci, významnější bude ten, který má nižší pořadí podle AltaVisty.

5 Výsledky

Odstranění neexistujících dokumentů. Jak již bylo uvedeno výše, průměrně 15% z AltaVistou nalezených dokumentů již neexistuje nebo je momentálně nedostupno. Takovéto dokumenty AVS odstraní. Počty dokumentů před a po odstranění neexistujících a momentálně nedostupných dokumentů ukazuje Tab. 1. Sloupec **Dotaz** obsahuje zadané dotazy, sloupec **AV** obsahuje počet požadovaných (nalezených) dokumentů pro zadaný dotaz, ve sloupci **Živé** je uveden počet odkazů po odstranění „mrtvých“ odkazů.

Dotaz	AV	Živé	Redukce (%)
+socket +program∗ perl source	10	9	10
decoder mp3 wav encoder	10	7	30
+python +tutorial	10	7	30
+spider tarantul∗ +picture gallery photo	10	7	30
Celkem	**40**	**30**	**25**
+linux +kde pictures	25	18	28
+history of +"computer science"	25	21	16
+spider +picture gallery	25	22	12
machine +learning logic	25	21	16
+popelinsky	25	21	16
"hand geometry"	25	20	20
"voice authentication"	25	20	20
"overlord system" NetLabs	25	18	28
+cactus picture gallery photo	25	23	8
+knowledge +discovery +databas∗	25	23	8
Celkem	**200**	**165**	**17,5**
+perl +program∗ +domain:cz	50	47	6
+tcl +tk description examples	50	40	20
linux xwindows +manager xf86	50	48	4
Celkem	**150**	**135**	**10**
Celkem pro všechny dotazy	**390**	**330**	**15,38**

Tabulka 1: Porovnání počtu nalezených a existujících odkazů

Poslední sloupec **Redukce (%)** uvádí v procentech poměr mezi počtem nalezených a výsledným počtem odkazů. Výsledkem tohoto kroku systmému AVS je zkvalitnění množiny nalezených dokumentů, která obsahuje pouze existující a dostupné dokumenty.

Přehlednější výstup. Zahlcení velkým množstvím odpovědí je pro uživatele AltaVisty poměrně nepříjemné, neboť ztěžuje hledání požadovaných dokumentů. Proto AVS počet nalezených dokumentů zredukuje vždy na jeden dokument z každého WWW serveru. Průměrná redukce nalezených odkazů je 32,5%.

Nevýznamný výskyt klíčového slova. Dokumenty, které neobsahují v čitelném textu ani jedno ze slov zadaných v dotazu, mají nulovou relevanci a tudíž budou ve výsledném seřazeném seznamu odkazů až úplně na konci — u AltaVisty tomu ovšem tak není. Tyto dokumenty lze jednoduše poznat podle údaje o počtu obsažených slov, jenž je u každého dokumentu uveden.

Řazení dokumentů. Objektivní posouzení toho, zda je jeden seznam dokumentů lépe seřazen podle relevance než druhý, je velmi obtížné. Proto také výsledky, které budou nyní prezentovány, jsou do jisté míry subjektivní. Jistá míra objektivnosti je zde však zachována, poněvadž na sběru výsledků se podílelo více uživatelů.

Na testování výsledné verze systému se podílelo celkem 5 osob. Testování bylo prováděno podle složitosti dotazů, kdy byly testovány náhodně zvolené dotazy jednoslovné, dvouslovné a dotazy složitější. Jednotlivé výsledky jsou shrnuty v tab. 2. Sloupec označený **AV** obsahuje počet odkazů nalezených AltaVistou, sloupec **AVS1** uvádí počet odkazů vypsaných systémem AVS. Sloupec **AVS2** obsahuje počet odkazů po odstranění všech neexistujících a momentálně nedostupných odkazů. Poslední nejvýznamnější sloupec tabulky označený **Výsledek** uvádí, který ze systémů (AVS, AV) dává podle uživatele lepší výsledky.

jednoslovné dotazy — V této kategorii získal převahu systém AVS, který v 60% případů vyprodukoval lepší výsledky než AltaVista, jejíž výsledky byly lepší pouze v 10% případů. Ve zbývajících 30% případů nebyla pozorována výraznější změna.

dvouslovné dotazy — V případě dvouslovných dotazů bylo výsledné pořadí dokumentů v polovině případů shodné. Pouze jeden z testovaných dotazů zpracovala lépe AltaVista (12,5%), zbývajících 37,5% dotazů bylo lépe zpracováno systémem AVS.

složitější dotazy — V kategorii složitých dotazů byl počet dotazů zpracovaných lépe systémem AVS shodný s počtem dotazů zpracovaných oběma systémy stejně — 44,5%. Zbývajících 11% případů zvládla lépe AltaVista.

Celkem tedy téměř v jedné polovině případů je pořadí dokumentů vrácené systémem AVS lepší než pořadí dokumentů určené AltaVistou a v 40,7% případů nebyla pozorována žádná výraznější změna. Pouze 11,1% dotazů přetřídila AltaVista lépe než AVS. Celkem téměř 89% dotazů bylo systémem AVS zpracováno stejně kvalitně nebo lépe než vyhledávačem AltaVista. Pouze v 11% případů byly výsledky AltaVisty lepší.

Lepší výsledky AVS lze přičíst následujícím skutečnostem. AVS pracuje pouze s nejaktuálnějšími verzemi dokumentů, naproti tomu AltaVista nezřídka operuje s dokumenty starými, neaktuálními. Kromě toho v AVS jsou ke třídění používány znalosti o struktuře dokumentů, vzájemných vztazích mezi jednotlivými částmi dokumentu a vztahu zadaných klíčových slov k relevanci dokumentu. Ne vždy je pořadí určené AltaVistou špatné. Proto také celkem v 40,7% dotazů nebylo pozorováno ani zlepšení ani zhoršení výsledků systému AVS v porovnání s výsledky AltaVisty. Horší výsledky AVS jsou způsobeny několika faktory.

Zadaný dotaz	AV	AVS1	AVS2	Výsledek
"expert system"	10	5	10	AVS
LZW	25	21	15	AVS
"data compression"	25	23	22	AVS
"Edgar Allan Poe"	25	24	16	AVS
"ballroom dancing"	10	9	9	stejné
kerberos	25	10	22	AVS
postscript	10	10	8	stejné
"hand geometry"	25	20	21	AVS
+organogermanium*	10	9	10	stejné
"Vaclav Havel"	50	10	48	AV
+spolecensky +tanec +domain:cz	25	19	20	stejne
+spider +tarantula	25	12	24	AVS
+"neural network" +backpropagation	25	24	24	stejné
+python +documentation	25	21	22	AVS
+"iron maiden" +discography	25	20	20	stejné
+"text compression" LZW	25	22	23	AVS
+"expert system" +"forward chaining"	10	10	9	AV
+richard +botik	25	12	20	stejné
+cactus +succu* +growing hint* how faq help	25	22	24	AVS
+history of +"computer science"	25	22	21	AV
+knowledge +discovery +databas*	50	36	41	stejné
+rabbit +wood +hunting forest	25	23	21	stejné
+ballroom +danc* photo gallery picture	50	34	45	AVS
+wine +sangria -champ* -restaurant* -bar	50	43	44	stejné
+search +"source code" +"site index" web	25	23	20	stejné
+tcl description tk language examp	25	20	19	AVS
+processor +intel +risc -i860	25	22	24	AVS

Tab. 2: Porovnání AltaVisty a AVS podle setříděných seznamů nalezených odkazů

V dokumentech byly používány jiné tagy (podobného významu) než ty tagy, které systém AVS považoval za významné. Navíc nelze ideálně nastavit váhy jednotlivých tagů a pravidel pro všechny možné případy.

6 Klady a zápory AVS

Mezi nalezenými dokumenty jsou pouze existující a dosažitelné dokumenty AVS pracuje pouze s nejaktuálnějšími verzemi dokumentů. Za velmi kladnou vlastnost považujeme skutečnost, že v mnoha případech (a to i tehdy, kdy výsledky AVS nejsou lepší než výsledky AltaVisty) najde mezi dokumenty, které jsou podle AltaVisty málo relevantní, některé dokumenty, které víceméně odpovídají zájmům uživatele a ty pak přemístí do předních částí výsledného seznamu. Seznam dokumentů, který AVS produkuje, je ve většině případů lepší nebo stejný jako seznam dokumentů vrácený vyhledávačem AltaVista

Viditelným nedostatkem systému je jeho relativní pomalost. Je to způsobeno tím, že všechny programy, které jsou při běhu systému spouštěny, běží přímo na WWW serveru, jenž je velmi přetěžován. Komunikační linky jsou značně vytížené a komunikace s AltaVistou resp. načítání samotných WWW stránek je tak značně zpomalována. Druhým důvodem načítání velkého množství dokumentů. Zrychlení celého systému by se dosáhlo provozováním systému na samostatném stroji, což bývá pro podobné systémy obvyklé. Pokud se však smíříme s dobou odezvy (nepřesahující při běžném provozu desitky minut), bude pro nás AVS velmi dobře použitelný.

AVS předpokládá jistý tvar výstupu AltaVisty. V případě jeho změny stačí změnit skript pro převod. Podobným způsobem by se řešilo případné použití jiného vyhledávače.

7 Další vývoj systému

Problematika určení relevance dokumentů je velmi široká a složitá. Systém AVS postihl pouze některé aspekty z této problematiky. Tudíž je ještě řada věcí, které mohou systém AVS přivést k produkci lepších výsledků. Uživatel by mohl mít možnost zadat doplňková klíčová slova, která nebudou použita pro vyhledávání, ale budou hrát důležitou roli při hledání optimálního pořadí nalezených dokumentů. Pořadí klíčových slov v dotazu může nést určitou informaci. Při kladení dotazů často člověk postupuje od konkrétních slov, která chce v dokumentu mít, až po slova, kterými doplňuje a upřesňuje své požadavky. Předchozí bod lze rozšířit tak, že uživatel ke každému klíčovému slovu zadá jakousi prioritu, podle níž jsou jím zadaná slova pro výsledné dokumenty důležitá.

Odkazy

1. Richard Botík: Inteligentní vyhledáváni ve WWW. Diplomová práce FI MU Brno, 1998.
2. V. Matyáš, L. Popelínský: *Vyhledávání na WWW*, Softwarové noviny č. 12/1996
3. Chris Culbert, Gary Riley, Brian Donnell: *CLIPS Reference Manual*