

Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky
Katedra informačního a znalostního inženýrství

Téma bakalářské práce:

Problémy vyhledávání informací
v Internetu – relevance,
důvěryhodnost a aktuálnost
vyhledaných informací

Vypracoval: Josef Kasík

Vedoucí práce: Ing. Vilém Sklenák, CSc.

Prohlášení:

Prohlašuji, že jsem předloženou bakalářskou práci zpracoval zcela samostatně a veškerou použitou literaturu a další podkladové materiály, které jsem použil, uvádím v bibliografii.

V Děčíně dne 3. ledna 2007

Josef Kasík

Poděkování

Děkuji Ing. Vilému Sklenákovi, CSc. a všem svým blízkým za čas, který mi věnovali, a za podnětné připomínky ke zlepšení mé bakalářské práce.

Abstrakt

Tématem této práce je rozbor tří stěžejních problémů z oblasti vyhledávání informací v Internetu. Jedná se o problémy relevance, důvěryhodnosti a aktuálnosti vyhledaných informací. Hlavním cílem práce je nejprve teoretický rozbor těchto problémů, obohacený o množství příkladů s nimi souvisejících. Tento rozbor je obsažen ve třetí kapitole práce, která následuje po dvou kapitolách úvodních. Předposlední kapitola je věnována analýze pěti vyhledávacích strojů, třech zahraničních a dvou čistě domácích. Ta je již prováděna z čistě praktického pohledu. Stejně tomu je u poslední kapitoly, kterou tvoří případová studie. Jejím jádrem je výzkum, provedený na šesti vybraných jednotlivcích, reprezentujících tři nejtypičtější skupiny uživatelů v oblasti vyhledávání informací v Internetu. Studie se především zaměřuje na použité metody vyhledávání a analyzuje též pokládané dotazy.

Úvod	- 3 -
1 Základní pojmy z oblasti informací	- 5 -
1.1 Co jsou data?	- 5 -
1.2 Co jsou informace?	- 5 -
1.3 Zaznamenávání informací	- 5 -
1.4 Kolik informací se vyprodukuje?	- 6 -
2 Vyhledávání informací v Internetu	- 8 -
2.1 Základní pojmy	- 8 -
2.2 Historie vyhledávání na Internetu	- 8 -
2.3 Proč k hledání informací využít právě Internetu?	- 8 -
2.4 Problémy vyhledávání informací v Internetu	- 9 -
3 Relevance, důvěryhodnost a aktuálnost vyhledávaných informací v Internetu	- 12 -
3.1 Relevance	- 12 -
3.1.1 Co je relevance?	- 12 -
3.1.2 Jak měřit relevanci?	- 12 -
3.1.3 Příčiny málo relevantních informací	- 15 -
3.1.4 Hledáme relevantnější informace, aneb jak správně vyhledávat?	- 18 -
3.1.5 Vyjádření speciálních požadavků na relevantní informace, aneb pokročilé možnosti vyhledávání	- 23 -
3.2 Důvěryhodnost	- 24 -
3.2.1 Co je důvěryhodnost a co důvěryhodnost webu?	- 24 -
3.2.2 Je Internet důvěryhodný?	- 24 -
3.2.3 Jak ohodnotit důvěryhodnost webu?	- 24 -
3.2.4 Jak učinit svůj web důvěryhodnějším?	- 27 -
3.2.5 Důvěryhodnost a vyhledávací stroje	- 28 -
3.3 Aktuálnost	- 28 -
3.3.1 Co je aktuálnost a jak na ni pohlížet?	- 28 -
3.3.2 Aktuálnost jako problém	- 29 -
3.3.3 Příčiny neaktuálních informací	- 29 -
3.3.4 Jak ovlivnit aktuálnost?	- 30 -
3.3.5 Aktuálnost a vyhledávací stroje	- 30 -
4 Vyhledávací stroje a jejich srovnání	- 32 -
4.1 Exalead	- 33 -
4.1.1 Obecné údaje	- 33 -
4.1.2 Konkrétní vlastnosti	- 33 -
4.1.3 Shrnutí	- 36 -
4.2 Google	- 36 -
4.2.1 Obecné údaje	- 36 -
4.2.2 Konkrétní vlastnosti	- 37 -
4.2.3 Shrnutí	- 40 -
4.3 Windows Live Search	- 41 -
4.3.1 Obecné údaje	- 41 -
4.3.2 Konkrétní vlastnosti	- 41 -
4.3.3 Shrnutí	- 43 -
4.4 Jyxo	- 43 -
4.4.1 Obecné údaje	- 43 -
4.4.2 Konkrétní vlastnosti	- 43 -
4.4.3 Shrnutí	- 45 -
4.5 Morfeo	- 46 -
4.5.1 Obecné údaje	- 46 -

4.5.2	Konkrétní vlastnosti.....	- 46 -
4.5.3	Shrnutí	- 48 -
5	Případová studie	- 49 -
5.1	Úvod	- 49 -
5.2	Průběh studie	- 49 -
5.3	Shrnutí studie.....	- 54 -
	Závěr.....	- 55 -
	Seznam použitých zdrojů	- 56 -
	Kapitola 1	- 56 -
	Kapitola 2	- 56 -
	Kapitola 3	- 57 -
	Kapitola 4	- 58 -

Úvod

Název bakalářské práce zní „Problémy vyhledávání informací v Internetu – relevance, důvěryhodnost a aktuálnost vyhledaných informací“. Problém vyhledávání informací obecně je bezpochyby aktuální v jakémkoli období. V dřívějších dobách bylo lidstvo odkázáno prakticky výlučně na tištěný materiál. S rozvojem moderních technologií a hlavně počítačů získal na vážnosti a důležitosti nový informační zdroj – Internet. Tvrdí se, že Internet zmenšuje svět. Jedním z předpokladů tohoto tvrzení je Internet jako volný zdroj informací z celého světa. Opravdu, pomocí Internetu nalezneme informace z druhého konce světa během několika vteřin a většinou zcela bezplatně. Toto tvrzení však skýtá jistá úskalí. Stinnou stránku Internetu je fakt, že se na něm vyskytují kvanta informací škodlivých a nepřesných, v některých případech až lživých. A právě proto je problém vyhledávání informací v Internetu zcela zásadní. Jak vyhledat právě to, co potřebuji? Jak zjistit, zda jsou vyhledané informace věcně i časově pravdivé? Na to se pokusím odpovědět v následující práci. Rozsah veškerých problémů při vyhledávání informací v Internetu však považuji za natolik obsáhlý, že jsem se rozhodl vybrat z tohoto bezpočtu tři konkrétní problémy. Výběr problémů byl zcela prozaický. Dle mého subjektivního názoru se jedná o tři nejzávažnější problémy spojené s vyhledávanými informacemi. Problémy, které v sobě ztělesňují většinu neduhů, jichž se uživatelé dopouštějí při procesu vyhledávání. Ačkoli o nich většina uživatelů ví, jsou pro ně informace získané z Internetu často absolutně směrodatné. A proto jsem téma bakalářské práce zvolil právě takové. Hlavním cílem práce je tedy rozbor problémů při vyhledávání informací v Internetu se zaměřením na tři výše zmíněné problémy.

Není v našich silách, abychom na Internetu sami sledovali veškeré nové informace. Proto k vyhledávání informací využíváme vyhledávacích služeb. Ve stále rostoucím prostředí World Wide Webu jsou vyhledávací stroje nezbytnými pomocníky pro efektivní přístup k informacím. Soubor uživatelů je však velice různorodý. Skládá se z počítačových nováčků na straně jedné a z vysoce kvalifikovaných odborníků na straně druhé. První kategorie většinou hledá informace z čistě „osobních“ důvodů, tj. aby se o tématu něco dozvěděla. Druhá kategorie však požaduje přesné a vysoce kvalitní informace, často pro profesionální využití. Profesionálové však mají ve vyhledávání informací trénink a zkušenosti. Ten počítačovým laikům nejen chybí, ale často ani nemají chuť se něčemu naučit. A tedy i vyhledávání informací v Internetu je pro profesionály snazší. V současné době je většina vyhledávacích strojů konstruována tak, aby vyhověla všem kategoriím uživatelů. To se sice snadno řekne, ale tento fakt klade na vyhledávací stroje ohromné požadavky. Ty by totiž měly vyhledat relevantní informace i tehdy, pokud byl uživatelův dotaz položen příliš vágně či nevhodně. Aby mohly vyhledávače uspokojit co nejširší škálu uživatelů, musí jejich tvůrci dokonale znát vyhledávací strategie všech typů uživatelů. Proto je zapotřebí chování uživatelů při vyhledávání informací neustále zkoumat.

Předchozí odstavec měl poukázat na fakt, že každý uživatel se od jiného nějakým způsobem liší. Stejně tak se něčím liší každé dva vyhledávací stroje. Liší se nejen podporovanými operátory, uživatelským rozhraním, ale také vyhledávacími algoritmy. Z toho zcela jasně vyplývá, že v každém vyhledávacím dostane uživatel poněkud odlišné výsledky svých dotazů. Jinými slovy, v každém vyhledávacím se budou nalezené výsledky lišit jak aktuálností a důvěryhodností, tak také relevancí. Proto jsem se rozhodl, že součástí práce bude též srovnání pěti vyhledávacích strojů, tři celosvětových a dvou českých. Ze zahraničních vyhledávačů jsem pro srovnání zvolil servery Exalead, Google a nový Windows Live Search. Domácí prostředí je zastoupeno vyhledávacími servery Morfeo a Jyxo. U každého vyhledávače zmíním nejen základní nástroje, ale také pokročilé metody vyhledávání a vlastnosti, kterými se liší od vyhledávačů ostatních. Celkové pojetí práce je směřováno k vyhledávání informací pomocí vyhledávacích strojů, takže předmětové katalogy zmíním jen okrajově. I rozbor problémů bude k tomuto faktu přihlížet.

Již jsem zmínil, že správné a účinné vyhledávání informací závisí též na lidském faktoru. Zadáte-li například vyhledání určitého pojmu starší osobě, s vysokou pravděpodobností sáhne

nejprve po nějakém výkladovém slovníku. Zadáte-li tentýž pojem osobě středního věku bez větších počítačových zkušeností, využije zřejmě služeb některého z vyhledávačů, avšak lhostejno kterého. A konečně zadáte-li stejný dotaz studentům či odborníkům z oblasti informatiky, využijí své zkušenosti s vyhledávací a tvorbou dotazů a naleznou požadovanou informaci poměrně rychle a efektivně. Z předchozích úvah plyne, že kvalitní vyhledávání informací záleží též na konkrétním uživateli, který dotaz pokládá. Již jsem předeslal důvody, které tvůrce vyhledávačů nutí analyzovat chování jejich „zákazníků“. I já jsem se tedy rozhodl provést případovou studii na šesti vybraných jednotlivcích, kteří svými znalostmi a zkušenostmi (popř. neznalostmi a nezkušenostmi) reprezentují šest nejčastějších typů dotazovatelů v oblasti Internetu. Škála dotazovatelů začíná od prostých laiků, pokračuje přes střední vrstvy počítačové inteligence a končí počítačově zdatnými jedinci. Případová studie by měla sloužit jako jakási sonda do mysli a způsobu uvažování těchto typů uživatelů při vyhledávání informací v Internetu. Osobně považuji případové studie a podobné výzkumy za velmi efektivní a vcelku jednoduchý způsob získávání informací o samotných uživateli. A protože mají sloužit vyhledávací stroje právě jim, považuji to za nezbytnou součást mé práce.

Úvod měl poukázat na souvislost problémů při vyhledávání informací se zdroji informací a typem dotazovatelů. Budu-li hledat informace o sluneční soustavě v Ottově slovníku naučném, bude mít vyhledaná informace nulovou aktuálnost ve srovnání s kterýmkoli internetovým zdrojem. Stejně tak hokejový fanoušek pátrající po údajích o počtu gólů Jaromíra Jágra v této sezóně bude mít naprosto přesnou informaci během dvaceti sekund, zatímco smyčcový virtuóz se jí nemusí dopátrat nikdy, protože zkrátka nebude vědět, kde hledat.

1 Základní pojmy z oblasti informací

1.1 Co jsou data?

Předtím, než přijde na řadu stěžejní pojem celé práce, je na místě uvést pojem data (jednotné číslo je datum). Budu se držet standardní definice, zmíněné např. v knize V. Sklenáka a kol. [1]: „**Data**...je označení pro čísla, text, zvuk, obraz, popř. jiné smyslové vjemy reprezentované v podobě vhodné pro zpracování počítačem.“ Data tvoří základ informací, neboť jsou nositeli informačního sdělení. Sama o sobě však nedávají žádný smysl.

1.2 Co jsou informace?

Pro definici slova informace tentokrát využiji internetové encyklopedie wikipedia [2]: „**Informace** je výsledkem zpracovávání, obsluhy a uspořádávání dat takovým způsobem, aby obohatila znalosti jejího příjemce.“ Srozumitelněji řečeno, informace jsou data, kterým je přiřazen význam či nějaký kontext. Z tohoto kontextu musí být zřejmé, co nám informace říkají. Informace pro nás tedy musí být srozumitelné. Ač si to mnozí z nás neuvědomují, s informacemi se setkáváme v drtivé většině každodenních činností. Každá činnost totiž vyžaduje informace. Pokud například člověk ráno vstane a jde z domova do práce či do školy, neobejde se bez několika závažných informací. Jestliže neví, co si obléci na sebe, pohlédne zpravidla ven na oblohu či na teploměr, aby zjistil, jaké je venku počasí. K tomu, aby se rozhodl, zda si obleče kalhoty či šortky, potřebuje zkrátka informace. Informace o tom, kolik je venku stupňů, zdali svítí slunce, zdali fouká vítr apod. Sdělení „venku je 25 stupňů“ je pro nás srozumitelná a smysluplná informace. Avšak samotnému číslu 25 bez jakéhokoli dalšího kontextu bychom jen stěží porozuměli. Je to pro nás jen nějaké abstraktní datum. Fakt, že si z informace o počtu stupňů venku dokážeme vyvodit nějaké závěry (např. to, že si oblečeme šortky), znamená, že máme určité znalosti. Zmíněný příklad zahrnoval pouze ty druhy informací, které jsou k dispozici bez větších problémů a zcela volně. V praxi se však nezdá, že setkáváme s takovými druhy informací, které musíme někde vyhledat, a to často z různých zdrojů. Někdy se bohužel stává, že ani nevíme (nemáme informaci o tom), v jakých zdrojích hledat. Důležitý je tedy způsob, jak a kde jsou informace zaznamenány a uloženy.

1.3 Zaznamenávání informací

Současnost nabízí v oblasti záznamu informací naprosto neskutečné možnosti, které si naši předkové ani neuměli představit. Jen okrajově zmíním prvotní záznamy informací v jeskynních malbách, na kamenných a hliněných destičkách a kůžích zvířat. V tu dobu se ještě skutečně nedalo hovořit o nějakém uceleném záznamu informací pro potřeby tehdejších lidí. Naprostý převrat v záznamu informací představoval vynález papíru ve starověkém Egyptě okolo roku 3 000 př. n. l. [3], který byl o 4,5 tisíciletí později přebit vynálezem knihtisku J. Gutenbergem (1456). Knihtisk drasticky snížil náklady na produkci knih, čímž umožnil jejich rozkvět. Ruční opisování knih bylo nespočetněkrát složitější než lehkost, s jakou tiskařské stroje po celé Evropě přinášely vědomosti stále širší vrstvě obyvatel. Nový impuls získal záznam informací v podstatě až koncem 19. století vynálezem fotografického filmu. Hlavní boom však přišel ve století dvacátém s nástupem moderních technologií – během něho spatřily světlo světa magnetofonové pásky, diskety, harddisky, kompaktní disky, minidisky, DVD, různé paměťové karty, flashdisky apod. V průběhu vývoje všech zmíněných typů záznamových médií přišel na svět Internet. Základy byly položeny již roku 1969 vznikem vojenské sítě ARPANET, která spojovala 4 významné uzly v USA [4]. O samotné historii Internetu byly napsány tisíce knih, a tak jen zmíním, že teprve v roce 1993 získal Internet současnou podobu a strukturu. To byl totiž rok, kdy do Internetu začaly mít přístup komerční firmy, nejprve z počítačové oblasti a později z oborů dalších. Nutno připomenout, že Internet sám o sobě není záznamovým médiem. Data jsou nadále fyzicky uložena na záznamových

médiích serverů, popř. domácích počítačů. Přesto se dá Internet bez přehánění nazvat nejobjemnějším záznamníkem informací na světě. Jedním z hlavních cílů veškerého zdokonalování a rozvíjení nosičů informací je jejich schopnost pomáhat při vyhledávání informací.

1.4 Kolik informací se vyprodukuje?

S volným přístupem komerčních subjektů do Internetu však neúměrně roste množství informací, které se na něm vyskytují. Než tedy přistoupím k samotnému vyhledávání informací, uvedu, v jak velkém prostoru informace vůbec vyhledáváme. Projekt výzkumných pracovníků pod vedením P. Lymana a H. R. Variana z Kalifornské univerzity v Berkeley již delší dobu zkoumá množství informací, které jsou každoročně vytvářeny, a to nejen na Internetu. Drtivá většina nově vytvořených informací je rozšiřována za pomoci 4 záznamových médií: magnetických, optických, papíru a filmu. Zbývající množství vytvářených informací je zanedbatelné. Takto uložené informace jsou šířeny čtyřmi informačními toky: telefonem, rádiem, televizí a Internetem. Poslední výzkum týmu pochází z roku 2003 [5] a pracuje s údaji z roku 2002. Jejich předchozí výzkum byl proveden v roce 2000 [6] a využíval údaje z roku 1999. Srovnáním těchto dvou výzkumů lze vypozorovat jeden alarmující trend.

Během oněch tří let rostl objem nově ukládaných informací tempem 30 % ročně. Za rok 2002 činil objem veškerých nově vyprodukovaných informací (na všech čtyřech médiích dohromady) cca 5 exabytů (EB), což je 5×2^{60} bytů. Přepočteno na jednoho obyvatele planety Země (při celkovém počtu 6,3 mld. obyvatel) to činí 800 megabytů informací na člověka za rok. Množství nově vyprodukovaných informací za rok 2002 znázorňuje tabulka 1 [7]. Horní odhad pracuje s velikostí dat po jejich potenciálním převedení do digitální podoby. Dolní odhad byl získán jeho předpokládanou kompresí. Celkem 92 % nově vytvořených informací bylo uloženo na magnetická média, převážně na harddisky.

Ukládací médium	Horní odhad v TB	Dolní odhad v TB
Papír	1 634	327
Film	420 254	76 690
Magnetická média	4 999 230	3 416 230
Optická média	103	51
CELKEM	5 421 221	3 416 281

tab. 1 - Množství nových informací za rok 2002

Pokud bychom převedli veškeré množství informací z americké knihovny Kongresu, která čítá na 17 milionů svazků, do datové podoby, dostali bychom se na číslo 136 terabytů (TB) informací. Oněch 5 exabytů nově vyprodukovaných informací je ekvivalentní 37 000 takových knihoven. Stejný výzkum nabízí i jiné přirovnání. 5 EB informací je suma rovnocenná všem slovům, které kdy lidstvo vyslovilo. Asi nikoho nepřekvapí, že jen samotné Spojené státy americké produkují okolo 40 % nově ukládaných informací. Čísla to jsou opravdu závratná.

Ale abych se příliš neodchýlil od tématu práce, vrátím se opět k samotnému Internetu. Stejný výzkum odhadl, že velikost Internetu byla v roce 2002 cca 533 000 terabytů informací. Z toho pouze 167 terabytů tvořil „viditelný“ nebo také „povrchový“ web. Jedná se o ten druh webu, u něhož většina uživatelů končí. Zahrnuje veškerý veřejně dostupný obsah, který je indexován vyhledávacími stroji. Pokud tedy vyhledáváme v Internetu pomocí vyhledávacích strojů, vyhledáváme v tomto prostoru. Oproti tomu web „neviditelný“ či „hluboký“ obsahoval 92 000 terabytů informací. Tento pojem zahrnuje především specializované dokumenty jako databáze, adresáře či dokumenty v jiných formátech (.pdf, .ps apod.). „Neviditelný“ web je hodnotnější jak z hlediska objemu informací (vždyť je 550krát objemnější), tak z hlediska obsahu pro uživatele

relevantních dokumentů. Dlužno podotknout, že drtivě největší množství přenášených dat v Internetu zahrnují emaily, v roce 2002 to bylo cca 441 000 terabytů. Zbylých 267 terabytů pokryly přenosy informací Instant Messagingu. Velikost Internetu znázorňuje tabulka 2 [8].

Oblast	Terabytů v roce 2002
Viditelný web	167
Neviditelný web	91 850
Emaily	440 606
Instant Messaging	274
CELKEM	532 897

tab. 2 - Velikost Internetu

V souvislosti s množstvím informací v Internetu zmíním pojem **informačního přehlcení**, který s tímto tématem velmi úzce souvisí. Opět využiji internetové encyklopedie wikipedia [9]: „Je to stav, kdy má uživatel příliš mnoho informací k tomu, aby učinil rozhodnutí či zůstal informován o určitém tématu.“ Jsou to tedy okolnosti, které nám ztěžují efektivní a úspěšné rozhodování. Důvodem je především skutečnost, že máme k dispozici nepřehledné množství informací, které nedokážeme zvládnout.

2 Vyhledávání informací v Internetu

2.1 Základní pojmy

Vyhledávání informací představuje určitý proces, jehož cílem je identifikace relevantních dokumentů nebo informací v informačních zdrojích [1]. Na jeho počátku je vždy přesně specifikovaná **informační potřeba** uživatele. Příkladem může být potřeba informací ohledně jízdních řádů vlaků. Tato potřeba je vyjádřena pomocí **informačního požadavku**, např. vyhledání vlaků z Brna do Prahy. Informační požadavek lze vyjádřit pomocí dotazovacího jazyka, výsledkem je pak **dotaz** [2]. Prvotním krokem při vyhledávání informací je vytipování vhodných **informačních zdrojů**, např. oficiální webové stránky Českých drah. Dalším krokem je samotné vyhledávání informací. Proces vyhledávání jako takový se označuje pojmem **rešerše** [3]. Lidé při vyhledávání informací využívají buď svých znalostí a zkušeností (pokud vědí co hledají a kde to mají hledat), nebo některé z vyhledávacích služeb (pokud mají o daném tématu chabé či vůbec žádné potuchy). Vyhledávání informací je vůbec nejčastějším důvodem použití Internetu ze strany uživatelů – dvě třetiny až tři čtvrtiny z nich používají Internet právě z tohoto důvodu [4].

Klíčovým slovem rozumíme slovo, kterým uživatel vyjadřuje svou informační potřebu jako součást dotazu [5]. Může jím být jednotlivé slovo, část slova nebo fráze. **Hit** (nebo také match) je dokument, nalezený vyhledávacím strojem a vyhovující zadanému dotazu [6]. **Index** je způsob organizace údajů, které jsou nashromážděny díky činnosti robota tak, aby bylo umožněno jejich rychlé a efektivní vyhledávání [7]. **Indexování** je tedy proces nalézání, třídění a ukládání údajů o umístění dokumentů.

2.2 Historie vyhledávání na Internetu

Vyhledávání v Internetu nemělo vždy stejnou podobu jako dnes, ale měnilo se spolu s rozvojem vyhledávacích služeb. Zcela prvním nástrojem pro vyhledávání informací v Internetu byla služba Archie z roku 1990 [8]. Sloužila k vyhledávání souborů na veřejných anonymních FTP serverech. Její hlavní nevýhodou byla neschopnost prohledávat obsahy souborů, nýbrž jen jejich umístění (tedy seznamy adresářů se soubory). V roce 1991 spatřil světlo světa Gopher, který však indexoval jen dokumenty obsahující jednoduchý text. Služeb Gopheru využíval vyhledávač Veronica (spuštěný v roce 1993), který prohledával prostory indexované právě touto službou. Jinak Gopher byl vůbec první službou svého druhu, jež umožňovala uživateli přistupovat k informačním zdrojům bez nutnosti znát jejich konkrétní umístění [9]. Prvním opravdu povedeným vyhledávacím strojem byl Aliweb z roku 1993, který běží dodnes. Vůbec nejstarším fulltextovým vyhledávačem (tedy vyhledávačem, umožňujícím hledat jakékoli slovo uvnitř webových stránek) byl WebCrawler z roku 1994. Stal se základem pro všechny pozdější vyhledávací služby a byl to také první veřejnosti opravdu známý vyhledávač. Totéž se dá říci o vyhledávači Lycos, spuštěném v témže roce, jehož hlavními konkurenty se staly AltaVista a Excite, fungující od roku 1995 [10]. Vzpomenul bych ještě prvního slavnějšího katalogového vyhledávače Yahoo!, který spustil činnost v dubnu 1994. Je to také jeden z mála vyhledávačů, jež se v konkurenci nevytratily a své postavení si drží dodnes. V roce 1998 byl uveden do provozu Google. O něm bude velmi podrobně pojednáno v kapitole zabývající se srovnáním vyhledávačů. Další ze srovnávaných vyhledávačů Exalead vznikl v roce 2004 [11] a Windows Live Search od Microsoftu je z roku 2006 [12], avšak vznikl syntézou se službou MSN, která na světě existuje již dva roky.

2.3 Proč k hledání informací využít právě Internetu?

Hlavní a bezkonkurenční výhodou Internetu jako informačního zdroje je jeho komplexnost. Jak praví jedna vžitá floskule: „Najdete tam všechno.“ Jenže není všechno jako všechno. Ano, najdete tam nikdy nekončící penzum informací ze všech oborů lidské činnosti – od kvantové

chemie, přes nejnovější módní trendy, výsledky sportovních zápasů konče. To je ono první „všechno“. Nenajdete tam ovšem adresy členů představenstva společnosti Eurotel či nákupní cenu piva, kterou účtuje pivovar Budvar svým odběratelům. Takové informace patří mezi ono druhé „všechno“. Nenajdete je tam z jednoho prostého důvodu – takové informace tam zkrátka nejsou. Ne každý si uvědomuje, že se na Internetu nevyskytuje naprosto vše, po čem se člověku zamane. Není to tedy zdroj univerzální, ale v drtivé většině případů znalosti nějakým způsobem obohatí.

Další obrovskou výhodou Internetu je rychlost vyhledávání. Zkuste si porovnat dobu, za kterou vyhledáte nějakou informaci v Internetu (mám tím na mysli čas mezi usednutím k počítači a nalezením první informace, která vás uspokojí) s dobou vyhledání téže informace v knize. Navíc pokud nemáte o dané oblasti příslušnou literaturu, pak bez toho, aniž byste opustili teplo domova, nemáte šanci nalézt informaci vůbec. Stejně je to u televize a rádia. Těžko lze předpokládat, že právě vysílají pořad o tom, co vás momentálně zajímá. U Internetu existuje jiné úskalí. Informaci na něm vyhledáte rychle pouze pod podmínkou, že víte, kde hledat. To ale platí u všech zdrojů. Rychlost strojové práce je každopádně větší než práce manuální.

Mezi nezanedbatelné výhody internetových informací patří jejich aktuálnost. On-line reportáže ze sportovních duelů, pohyby cen akcií či měnových kurzů, aktuální počasí – nikde jinde nenajdete tyto informace čerstvější než na Internetu. Snadnost a pohodlnost vyhledávání jsou dalšími přednostmi Internetu v porovnání s jinými informačními zdroji. Naneštěstí existuje dost lidí, kteří onu pohodlnost vyhledávání v Internetu preferují před kvalitou získaných informací. Další výhody uvedu pouze výčtem: více uživatelů může pracovat se stejnými informacemi najednou, dotazy lze pokládat v jazyce blízkém přirozenému jazyku, vyhledávat můžeme kdykoli a jak dlouho chceme.

Jistě jste si všimli, že téměř u každé výhody jsem obratem uvedl s ní související nevýhodu. Je to dáno tím, že největší výhody vyhledávání v Internetu jsou také jeho velkými slabiny. Vyhledávání v Internetu nebude rychlé, pokud nebudeme vědět, kde hledat. Tak jako je na webu spousta aktuálních informací, je tam i spousta zastaralých informací. Podlehne-li iluzi, že je na Internetu vše, pak se na něm budeme pít po sebemenších maličkostech, aniž bychom měli šanci je najít. Pokud bychom vše hledali z domova, zlenivíme. Zkrátka a dobře musíme vědět, kde informace hledat, jak je hledat a jakým informacím můžeme důvěřovat. Často je též dobré položit si otázku, zda je vůbec účelné užít k nalezení informací Internetu.

2.4 Problémy vyhledávání informací v Internetu

Konečně se dostávám k samotnému jádru práce. V 1. kapitole jsem zmínil pojem „viditelného“ webu jako prostoru, který je indexován vyhledávacími stroji. Je tedy poněkud zvláštní, že existují dokumenty, které se vyskytují na Internetu a přesto je nelze nalézt pomocí vyhledávacích strojů. Jedná se především o dokumenty, které jsou k dispozici na vyžádání prostřednictvím vyhledávacího formuláře, dokumenty chráněné heslem a dokumenty označené jako výjimka pro roboty [13]. Uvedu 2 příklady za všechny. Při nakupování zboží na Internetu se často po vyplnění formulářů souvisejících s kontakty objeví stránka, vyzývající k potvrzení zadaných údajů. Tuto stránku žádný vyhledávací stroj nenajde. Stejně je tomu u dokumentů v naší emailové schránce, které jsou pro změnu chráněny naším heslem. Velikost Internetu je tedy obrovská, avšak vyhledávací stroje ji v žádném případě nemohou pokrýt celou. Pak se ale může stát, že pro nás ideální dokument se na Internetu sice vyskytuje, my však nemáme šanci ho nalézt. Celkové množství dokumentů na Internetu se odhaduje na 11,5 miliard (leden 2005) [14]. Google, vyhledávač s největším indexem vůbec (8 miliard), zajišťuje pokrytí tohoto množství z cca 70 %.

Většina problémů spojených s vyhledáváním informací v Internetu pramení z jeho samotné podstaty. Internet je otevřený, přístupný a dosti svobodný. Kdokoli na něj může umístit prakticky cokoli. Neexistuje zde žádná cenzura. Samo sebou též nikdo nemůže nikoho donutit, aby své stránky pravidelně aktualizoval. Velmi výstižně charakterizuje prostředí Internetu následující věta [15]: „Internet je nejen monumentálním zdrojem informací, ale též dezinformací.“ Dokonce se na

něm vyskytují kvanta lživých informací, což ovšem není nic, co by Internet odlišovalo od jiných informačních zdrojů.

Dva problémy jsou podle mne nejzávažnější. Jsou jimi správná volba vyhledávacího systému a vhodná konstrukce dotazu pro něj. Jejich řešení je bohužel do značné míry záležitostí zkušeností. Jednotlivé vyhledávací služby se určitým způsobem liší. Různé služby tak mohou i přes stejný dotaz přinést jiné výsledky. Uživatel tedy musí rozmyslet, jakou z nich použít v závislosti na povaze informační potřeby. Obecně lze říci, že neexistuje jeden vyhledávací systém, který by byl vhodný pro všechny typy požadavků. Kdyby takový existoval, nebylo by ostatně potřeba těch dalších. Vyhledávacími stroji se podrobně zabývá celá 4. kapitola, která rovněž pomůže určit, jaký vyhledávač užít a kdy ho užít.

Správné položení dotazu je kapitola sama pro sebe a v této práci se tím také v nemalém měřítku zabývám. Na něm je totiž nalezení relevantních informací značně závislé. Správný dotaz určuje množství faktorů. Dotaz by neměl obsahovat překlepy a pravopisné chyby, neměl by být příliš obecný ani konkrétní, klíčových slov by nemělo být ani moc ani málo apod. Uměním není najít co nejvíce výsledků, ale najít co nejpřesnější výsledky. Najít miliony odkazů není ani záslužné, ani výhodné.

Samotná povaha uživatelů je další příčinou problémů při nalézání informací. Uživatelé jsou často netrpěliví a potřebují výsledky rychle. Jsou dosti pohodlní na to, aby si o zásadách vyhledávání našli byť jen ty nejzákladnější poučky. Takový je naneštěstí profil nejběžnějšího uživatele. Pokud jim tedy vyhledávací stroj odpoví „nebyl nalezen žádný odkaz“, pak v lepším případě svůj dotaz upraví, v horším případě vyhledávání vzdají. Navíc velká část jedinců z první kategorie se často vzdá při druhém neúspěchu. Na opačném konci je situace, kdy uživatel najde množství informací právě o tom, co potřeboval, avšak sahá pouze po jedné z nich. Žádnou kontrolu pravdivosti ani srovnání s jinými výsledky neprovádí. Důvěřuje zkrátka jedné jediné informaci. Tím se ovšem vystavuje netušenému riziku. Riziku, že informace, kterou vybral, nebude důvěryhodná. Navíc musí dát uživatel pozor, aby bez uvedení referencí doslovně nekopíroval text, který je pod autorským zákonem, což je ovšem stejné u všech zdrojů. S lidským faktorem rovněž souvisí umění učit se z vlastních chyb a sbírat zkušenosti z každého vyhledávání (ať již úspěšného či neúspěšného). S jejich pomocí lze k podobným informačním potřebám klást podobné dotazy. Nebo například zjistím, že mi jistý vyhledávač poskytl informaci, z jejíhož využití mi plynul určitý neprospěch. Pak zkrátka již jeho služeb nevyužiji.

Mezi důležité problémy při vyhledávání informací patří také nedostupnost vyhledaných informací. Každému se jistě stalo, že mu vyhledávač sice našel odkaz, který však již nebyl v okamžiku hledání dostupný (viz známá hláška „server not found“). Stává se to v situacích, kdy vyhledávací stroj sice zaindexoval určitou stránku (resp. její umístění), ale data na ní již nejsou fyzicky k dispozici. Data byla v periodě po poslední indexaci smazána a nová indexace, která by to zaznamenala, ještě neproběhla. Vyhledány jsou tedy neaktuální odkazy. Vadí nám to především v situacích, kdy si pro pozdější použití uložíme pouze odkaz na nějakou stránku. Při pozdějším vyhledání stránky již požadované informace nejsou k dispozici. Opětovné hledání přemístěných dat je velmi nepohodlné.

Málo z nás si uvědomuje, že informace nejsou jen něco abstraktního. Něco, co vždy získáme zcela zadarmo a tak se k tomu také chováme. Informace jsou také zbožím [16] a jako s každým zbožím se s nimi dá velmi snadno obchodovat. Z tohoto pojetí vyplývá, že za vyhledání či poskytnutí určitých druhů informací se také musí řádně zaplatit. Co to však musí být za informace, že je za ně člověk ochoten zaplatit? Může se jednat o výpisy z trestního rejstříku či katastrálního úřadu, rady ohledně nákupu akcií, nebo také archivy důležitých dokumentů. Řada podobných informací je dostupná přes tzv. zdroje komerčního typu, jako jsou např. databázová centra. Spousta lidí však neplatí ani za mnohem důležitější a samozřejmější věci, než jsou informace. Přijde jim to zkrátka jako vyhazování peněz. A přitom informace z podobných databází mají zaručenou kvalitu.

Záleží jen na nás a našem zájmu (popř. nezájmu) o kvalitní a zaručené informace. Využívání komerčních zdrojů také ovlivňuje fakt, že uživatelé o jejich existenci často ani nevědí.

Shrňme si tedy **nejzávažnější problémy vyhledávání informací v Internetu**, které způsobují, že uživatel nenajde to, co potřeboval, nebo že vybere informaci nějakým způsobem falešnou:

- neúplné pokrytí Internetu vyhledávacími službami
- nesprávně položený dotaz
- nesprávná volba vyhledávacího systému
- lidský faktor – netrpělivost, zbrklost, pohodlnost, malé zkušenosti s vyhledáváním, spoléhání se na neověřené informace
- pozor na autorská práva použitých informací
- nedostupnost odkazů, jejichž obsah byl smazán či přemístěn
- nevyužití informačních zdrojů komerčního typu

Zmíněné problémy jsou bezpochyby neopominutelné. Ale všechny jsou jen konkrétním vyjádřením problémů o něco obecnějších. Některé z nich jsem již zmínil v předešlých částech práce. Jsou to ona tři magická slova – relevance, důvěryhodnost, aktuálnost. Pokud nepoložím správně dotaz a nezvolím vhodný vyhledávací systém, zobrazené výsledky nebudou dosti relevantní. Pokud nebudu trpělivý, nenajdu nic. Neověřím-li informaci z více odkazů, vystavuji se riziku nedůvěryhodnosti této informace. Nedostupnost odkazů svědčí o jejich neaktuálnosti. Nevyužitím komerčních informačních zdrojů budu ochuzen o velké množství 100 % důvěryhodných informací. Takto bychom mohli pokračovat do nekonečna.

3 Relevance, důvěryhodnost a aktuálnost vyhledávaných informací v Internetu

Tyto tři pojmy souvisí především s hodnocením kvality získaných informací. U většiny informací požadujeme jejich maximalizaci. Pokud ovšem záměrně hledáme starší informace (např. v archivech článků), požadavek maximální aktuálnosti samozřejmě vypouštíme. Problém relevance je z těchto tří nejzásadnější, poněvadž se týká činností, které samotnému vyhledávání teprve předcházejí. Navíc relevanci informací, které budou vyhledány, může uživatel do značné míry ovlivnit. To se týká hlavně správného položení dotazu a výběru vhodného vyhledávacího systému. Důvěryhodnost a aktuálnost souvisí spíše se samotným výběrem správných informací z množství nalezených výsledků.

3.1 Relevance

3.1.1 Co je relevance?

Relevance obecně je charakterizována jako důležitost, závažnost, podstatnost [1]. V oblasti získávání informací se pojem **relevance** chápe jako vlastnost vztahu mezi dotazem uživatele a jednotlivým dokumentem jako prvkem množiny všech nalezených dokumentů [2]. V užším smyslu je to číselné ohodnocení výsledku vyhledávání, které představuje míru toho, jak moc výsledek odpovídá informační potřebě vyjádřené dotazem [3]. Lze na ni pohlížet jako na informativnost dokumentu. Informace je tedy relevantní, pokud uspokojuje informační potřebu uživatele. Přináší-li mu něco nového, pak je relevantní. Relevanci lze chápat dvojím způsobem. Vyhledávací stroje ji chápou jako podobnost mezi zadaným dotazem a vyhledávacími obrazy dokumentů. V subjektivním hledisku je chápána jako rozdíl mezi dvěma názory uživatelů se stejnou informační potřebou na tytéž výsledky. Co je relevantní pro jednoho uživatele, nemusí být relevantní pro toho druhého. Takto subjektivně chápáná relevance se označuje jako **pertinence** [4].

3.1.2 Jak měřit relevanci?

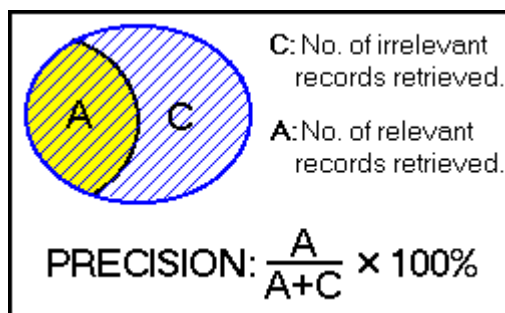
Nyní se dostávám k nejožehavějšímu úkolu spojenému s relevancí. Tou je její měřitelnost. Subjektivní relevance je měřena z pohledu uživatele a je tudíž u každého jiná. Objektivní relevanci, tj. vztah mezi dotazem uživatele a vyhledávanými dokumenty, musí řešit příslušný vyhledávací systém. Ať již má v sobě systém zabudovanou jakoukoli metodu posouzení relevance výsledků, jeho úspěšnost lze měřit dvěma stěžejními charakteristikami – přesností a úplností. To jsou též nejpoužívanější míry hodnocení efektivnosti vyhledávání.

Každý z dokumentů ve výsledcích má pro uživatele jinou míru relevance vzhledem k jeho informační potřebě (každý má tedy jinou subjektivní relevanci). Pro měření však předpokládáme, že pro daný dotaz lze všechny hity ve výsledcích rozdělit na relevantní a irrelevantní, tj. v úvahu nelze brát přesnou míru relevance pro uživatele. Označme dotaz písmenem q (q je zde jen indexem), množinu všech nalezených dokumentů NAL_q a množinu všech relevantních dokumentů REL_q . Potom **přesnost (precision) P_q** definujeme jako [5]:

$$P_q = \frac{|REL_q \cap NAL_q|}{|NAL_q|}$$

Přesnost vyjadřuje, jak velká část nalezených dokumentů je relevantní. Početně se tedy jedná o poměr relevantních nálezů k celkovému počtu nálezů. V nějaké obrovské databázi (např. databázi webových stran indexovaných Googlem) mějme například 1000 webových stran o Jaromíru Jágrovi. Vyhledávač nalezne v celé databázi 100 odkazů, z nichž 50 je pro uživatele relevantních.

Pak přesnost vyhledávání činí 0,50. Ze vzorce vyplývá, že přesnost se pohybuje v intervalu $<0,1>$. Její grafické znázornění obsahuje obrázek 1 [6]. Přesnost je vyjádřena jako podíl žluté plochy na celkové ploše oválu. Pro procentuální vyjádření musíme přesnost vynásobit stem.

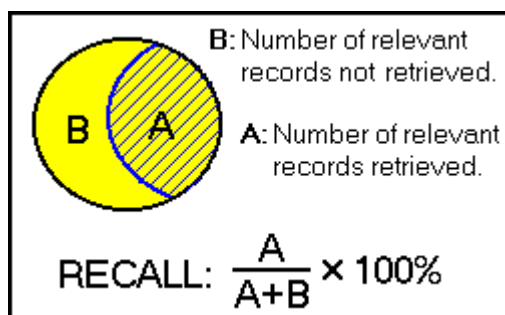


obr. 1 - Znázornění přesnosti

Úplností (recall) R_q rozumíme [7]:

$$R_q = \frac{|\text{REL}_q \cap \text{NAL}_q|}{|\text{REL}_q|}$$

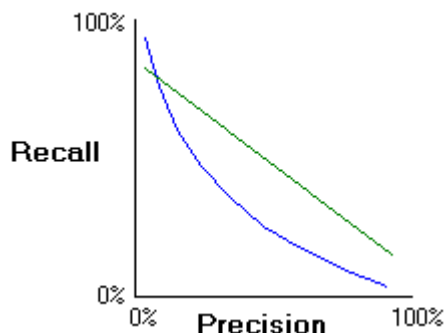
Úplnost nás informuje o tom, jak velká část relevantních dokumentů se vůbec našla. Početně je to tedy poměr získaných relevantních nálezů k celkovému počtu relevantních záznamů v databázi. Použiji stejný příklad jako u přesnosti. Pomocí dotazu uživatele bylo nalezeno 100 internetových stránek o Jaromíru Jágrovi. Úplnost tedy činí 0,10 (našlo se 100 relevantních odkazů z celkem 1000 možných). I zde je zřejmé, že úplnost patří vždy do intervalu $<0,1>$. Grafické znázornění úplnosti obsahuje obrázek 2 [8]. Součet $A + B$ značí celkový počet relevantních hitů v databázi. Plocha A (ta vyšrafovaná) představuje počet relevantních dokumentů, které byly nalezeny. Plocha B představuje relevantní dokumenty, které nebyly nalezeny. Úplnost je zde znázorněna jako podíl vyšrafované plochy (A) na celkové ploše oválu ($A + B$). Opět platí, že pro procentuální vyjádření musíme úplnost vynásobit stem.



obr. 2 - Znázornění úplnosti

Zcela jistě bychom byli rádi, kdyby vyhledávací metoda maximalizovala oba ukazatele. Ideálním případem by bylo $P_q = R_q = 1$. To by znamenalo, že všechny relevantní dokumenty z databáze byly nalezeny a že výstup vyhledávání neobsahuje žádné jiné dokumenty, než právě ty relevantní. V praxi se však ukazuje, že tyto veličiny jsou na sobě nepřímo závislé. Důvodem nepřímé závislosti je samotný přirozený jazyk. Pokud bychom totiž chtěli získat vyčerpávající (tedy úplné) množství informací o daném tématu, pak musíme brát v potaz veškerá synonyma, související výrazy, širší a obecnější výrazy apod. Následkem toho však zákonitě utrpí přesnost. Význam synonym nemusí být vždy naprosto stejný jako význam původního slova. Z toho důvodu roste pravděpodobnost nalezení irrelevantních záznamů. Stejně tak použití širších termínů nás může

přivést k materiálu, který se jen velmi okrajově dotýká našeho ohniska zájmu. Inverzní vztah přesnosti a úplnosti ilustruje obrázek 3 [9]. Modrá křivka značí úplnost, zelená přesnost. Je vidět, že s rostoucí úplností klesá přesnost a naopak. Přesný sklon a tvar obou křivek se sice liší v závislosti na vyhledávacím systému, avšak obecný vztah inverze je u všech systémů stejný.



obr. 3 - Vztah přesnosti a úplnosti

Při vyhledávání je nutno brát v potaz obě veličiny. Velmi častý je chybný úsudek uživatele, kdy je pro něj rozhodující počet nalezených dokumentů, tudíž bere v úvahu pouze úplnost a ne přesnost. Vyšší přesnost je však často důležitější než úplnost. Už jen proto, že nemusíme zkoumat tak velké množství dokumentů, u nichž není zaručena jejich relevance. Zkoumáme spíše malé množství více relevantních dokumentů. Relevantní hity poznáme ve výsledcích vyhledávání velice snadno. Většina vyhledávacích systémů totiž standardně řadí výsledky právě podle relevance. Ostatně pokud chceme uspokojit náš informační požadavek, je toto řazení jediné smysluplné.

Přesnost a úplnost rovněž závisí na povaze našich požadavků. Hledáme-li odpověď na nějakou konkrétní otázku, pak je pro nás velice důležitá přesnost výsledků. Té dosáhneme jedině precizně pokládanými dotazy, složenými často z více slov. Úplnost pro nás není stěžejní. Pokud naopak hledáme pouze základní a zevrubné informace, není pro nás důležitá ani přesnost, ani úplnost. A konečně hledáme-li co nejvíce dokumentů na dané téma, požadujeme maximální úplnost. Tím však logicky trpí přesnost. Větší úplnosti dosáhneme pokládáním širokých dotazů.

Získané hodnoty přesnosti a úplnosti však skýtají i určité obtíže. Jak jsem již zmínil, posouzení relevance a irelevance je záležitostí čistě subjektivní. Zřídka kdy je dokument 100 % relevantní či 100 % irelevantní. Navíc samotné měření není vždy zcela jednoduché. To se týká hlavně úplnosti. Často je problematické stanovit, kolik relevantních záznamů existuje v celé databázi (tj. kolik je REL_q). Úplnost se tak musí odhadovat. K jejímu určení se používají různé metody. Například jedna z nich pracuje s relevantními záznamy nalezenými v rozdílných vyhledáváních. Každá vyhledávací metoda musí vždy řešit tři rozhodnutí – jak vyjádřit dotaz, jak dokument a kdy řekneme, že dokument je pro jistý odkaz relevantní. V závislosti na těchto třech otázkách se konstruuji vyhledávací modely. Jejich tvorba však překračuje rámec této práce.

Objektivní relevance, tj. podobnost mezi dotazem uživatele a vyhledávacími obrazy dokumentů, je také důležitým faktorem. Konkrétní výpočty relevance u jednotlivých vyhledávačů jsou vcelku logicky utajovány. Základní a obecné principy výpočtů jsou však společné a veřejnosti známé [10]. Dotaz vyhledávacímu systému pokládáme ve formě klíčových slov. Ta jsou také důležitá při jeho vyhodnocování, kdy vyhledávač porovnává klíčová slova dotazu se slovy v:

- názvu dokumentu – jeho základní charakteristika zobrazená v záhlaví prohlížeče, ve zdrojovém kódu HTML se jedná o element „title“
- popisu dokumentu a klíčových slovech – stručný popis dokumentu a jeho klíčových slov, zahrnutý jako tzv. metainformace, ve zdrojovém kódu HTML se jedná o metataggy „description“ a „keywords“, tyto informace uživatel při prohlížení dokumentu nevidí, jsou umístěny v hlavičce, jsou však důležité při indexování a prezentaci dokumentu jako hitu při vyhledávání

- těle dokumentu – vlastní obsah dokumentu, který se zobrazuje uživateli

Při procesu vyhodnocování dotazu jsou nacházené dokumenty řazeny sestupně dle relevance, která je vypočítávána pomocí speciálních algoritmů, typických pro každý vyhledávač. Na pořadí dokumentů ve výsledcích má vliv množství faktorů. Nejdůležitějšími z nich jsou pořadí klíčového slova v dokumentu (čím blíže k počátku, tím výše ve výsledcích), frekvence klíčového slova (čím větší frekvence, tím výše), výskyt klíčového slova v názvu dokumentu a nadpisech (důležitější jsou slova v názvu či metadatech a nadpisech vyšší úrovně) a blízkost klíčových slov (relevantnější jsou dokumenty s klíčovými slovy blízko sebe).

Stinnou stránkou těchto v zásadě jednoduchých metod určení relevance je jejich snadné zneužití. Lidé totiž vědí, jak tyto metody fungují. Při rešerši se vždy jedná o porovnávání klíčových slov v dotazu s jejich výskytem v indexovaných dokumentech. Pokud tedy někdo chce mít své stránky v popředí vyhledaných výsledků, může do svých dokumentů vhodně rozmístit klíčová slova. A tak se stává, že se na předních místech nalezených hitů objevují dokumenty, které s požadovaným tématem souvisí jen okrajově. A jak uvidíme v další podkapitole, co je v popředí, to se „počítá“.

V souvislosti s tímto problémem ještě uvedu způsob určení relevance dokumentu, kterého využívá vyhledávač Google [11]. Tím je algoritmus **PageRank**, vyvinutý Larry Pagem a Sergeyem Brinnem. Využívá vzájemné provázanosti hypertextových odkazů, kdy nějaká stránka „doporučuje“ jinou. V úvahu však rovněž bere samotné hodnocení původní odkazující stránky. Z těchto údajů pak algoritmus vypočítává určité číslo, které se označuje jako PageRank. Čím vyšší PageRank stránka má, tím lépe pro ni. Tím lépe ve smyslu jejího umístění jako hitu v nalezených výsledcích. PageRank je ve své podstatě velmi jednoduchý algoritmus. V jednoduchosti je sice krása, ale také snadnost zneužití. A tak se i algoritmus PageRank stal snadným cílem. Jeho vlastností se zneužívá k tvorbě tzv. Google bomb [12]. Jejich podstata je jednoduchá. Někdo na své (nebo ještě lépe na několika svých) webové stránce (stránkách) umístí velké množství odkazů na jistou cílovou stránku a odkazy pojmenuje nějakým urážlivým textem. Cílová stránka tak získá vysoký PageRank a Google na ni bude po zadání dotazu v podobě onoho urážlivě laděného textu odkazovat. Je-li odkazů mnoho, může se cílová stránka objevit i na prvním místě ve výsledcích, což je obvykle také cílem Google bomb. Smyslem těchto bomb je většinou diskreditace osob neoblíbených u autora bomby, nebo zkrátka pokus o zviditelnění své stránky. První veřejně známá Google bomba pochází z roku 1999. Tehdy mohli uživatelé nalézt na prvním místě s výsledky po zadání dotazu *more evil than Satan* stránky firmy Microsoft. Od té doby se stali oběťmi Google bomb mnohé známé osobnosti, např. G. W. Bush, S. Berlusconi, M. Grebeníček a další. V březnu letošního roku okusil hořkost Google bomby také Jiří Paroubek, na jehož stránky stále ještě odkazuje dotaz *namyšlenej buran*.

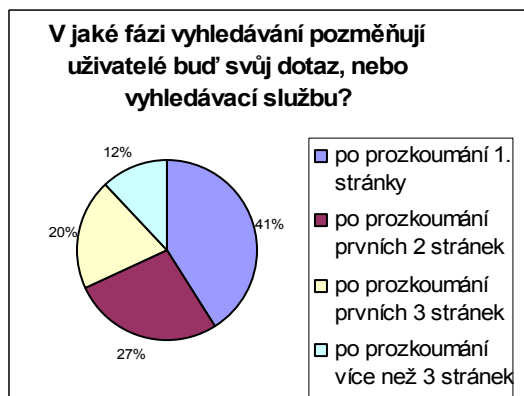
3.1.3 Příčiny málo relevantních informací

Jednou z hlavních příčin nesprávného vyhledávání obecně je samotný uživatel. Především na něm závisí úspěch při vyhledávání informací. Proto je právě chování uživatele při vyhledávání předmětem zkoumání řady studií. Z těchto studií bohužel nevychází uživatelé příliš lichotivě. Jeden z výzkumů byl proveden na počátku roku 2006 firmou iProspect, která se výzkumem vyhledávacích strojů zabývá [13]. Dovolte mi tedy několik údajů, které osvětlí pár velmi zajímavých skutečností.

Prvním alarmujícím číslem je 62 % uživatelů, kteří kliknou na odkaz hned na první stránce s výsledky. Plných 90 % uživatelů končí nejpозději na třetí stránce (viz obrázek 4 na následující straně). Pokud tedy vlastníte firmu, snažte se zajistit, aby odkazy na vaše webové stránky byly na prvních třech stránkách s výsledky. Jinak se totiž může stát, že vás na Internetu nikdo nenajde (pokud tedy nezná vaši adresu přímo). Celkem 41 % uživatelů pozměňuje svůj dotaz či vyhledávací službu, pokud nenajde nic relevantního na první stránce s výsledky. Celých 88 % se k tomuto kroku uchýlí po neúspěchu na prvních třech stranách (viz obrázek 5 na následující straně).



obr. 4



obr. 5

Pokud je první vyhledání neúspěšné, pak 82 % uživatelů specifikuje svůj původní dotaz dalším klíčovým slovem (popř. slovy), přičemž vyhledávací služba zůstává zachována. Tato skutečnost prokazuje určitý stupeň loajality uživatelů ke své vyhledávací službě. Určitě jí věří více než svému dotazu. Zde lze vypožorovat další impuls pro firmy, aby do zdrojových kódů svých webových stránek uváděly i delší a specifičtější klíčová slova. Poslední statistika se týká ponětí uživatelů o významu pořadí nalezených výsledků. Celkem 36 % uživatelů si myslí, že společnosti, jejichž webové stránky jsou na předních místech mezi nalezenými výsledky, jsou také špičky v daném oboru. Plných 39 % se k tomuto tvrzení staví neutrálně a pouze 25 % uživatelů si myslí, že tomu tak není. Z toho očividně plyne mylná představa části uživatelů, že vyhledávací stroj řadí výsledky podle vedoucího postavení firem v daném odvětví. Marketingoví odborníci velkých firem o těchto mylných úsudcích samozřejmě vědí, a tak vynakládají značné úsilí, aby webové stránky jejich firmy byly minimálně mezi prvními třiceti nalezenými výsledky.

Ze studie plyne první klíčová skutečnost. Drtivá většina uživatelů končí procházení odkazů nejpozději na 3. stránce s výsledky. Co když se však relevantní dokumenty vyskytují na stránkách dalších? Pak mají uživatelé zkrátka smůlu. Dalším problémem je skutečnost, že jen málo uživatelů skutečně ví, podle čeho jejich vyhledávač řadí nalezené výsledky. Zkrátka si pro sebe nedokáží interpretovat to, co našli.

Velkým problémem je též struktura dotazu uživatelů. Mnohým z nich totiž zcela uniká základní vyhledávací metodika. Pokud jde člověk do knihkupectví či obchodu, většinou dokáže vysvětlit, co hledá. K dispozici má totiž přirozený, a tudíž bohatý jazyk. Při vyhledávání informací však uživatel nemůže žádným způsobem vyhledávači vysvětlit, co přesně hledá, jaké to má mít parametry apod. Musí si vystačit s určitým dotazovacím jazykem. A právě zde je problém. Mnozí uživatelé nevidí mezi přirozeným a dotazovacím jazykem žádný rozdíl. Pravdou je, že v současné době se tvůrci vyhledávacích systémů snaží dosáhnout toho, aby mezi nimi skutečně žádný rozdíl nebyl. N. J. Belkin ze Státní univerzity v New Jersey shrnuje problémy uživatelů velmi výstižně [14]: „Jak uhádnout slova, která je třeba použít v dotazu tak, aby adekvátně vystihla problém uživatele a zároveň byla totožná se slovy, které vyhledávací systém používá k jeho vyjádření?“ Důsledkem této myšlenky je fakt, že jedno téma může být indexátorem a uživatelem chápáno zcela rozdílně.

Studie prokazují [15], že mezi tři hlavní faktory ovlivňující formulaci dotazu patří odborná znalost prostředků vyhledávání (tj. vyhledávacího systému, počítačů obecně apod.), odborná znalost vyhledávaného okruhu a typ vyhledávání (otevřené vs. uzavřené otázky). Obecně lze říci, že průměrný uživatel používá kratší dotazy (obvykle 1–3 slova) a pokročilé operátory (viz dále) používá velice zřídka. Je taktéž prokázáno, že chování expertů a nováčků při vyhledávání se výrazně liší. Při formulaci dotazu používá expert obvykle delšího (tudíž komplexnějšího) dotazu než méně znalý jedinec. Rovněž pokročilé operátory mu nejsou cizí. Experti se snaží dostat přímo

k informacím, poněvadž vědí, jak položit ten správný a konkrétní dotaz. Méně zkušení uživatelé většinou volí značně obecný a široký počáteční dotaz. Chtějí, aby jim vyhledávací systém našel alespoň nějaké vodící odkazy, ze kterých by se dalo navigací dostat k něčemu konkrétnějšímu. Rozdíly jsou však i ve vyhledávací strategii. Počítačový expert ji dokáže flexibilně měnit, zatímco strategie technických amatérů (pokud vůbec nějakou mají) je značně rigidní. Stejně jsou rozdíly mezi uživateli znalými vyhledávaného tématu a laiky v daném oboru.

Z hlediska relevance je velký problém v užívání širokých dotazů. Za **široký se považuje dotaz**, v němž figurují slova obecnější, než slova v názvu „úlohy“. Úlohou mám na mysli vyjádření informačního požadavku buď konkrétní otázkou, nebo nějakým obecnějším názvem tématu. Pokud se v dotazu vyskytuje polovina nebo méně než polovina klíčových výrazů z víceslovného vyjádření hledané „úlohy“, pak je dotaz považován za široký [16]. Hledáme-li např. odpověď na otázku „Jak se jmenuje předseda dolní komory parlamentu Ruské federace?“, pak dotaz *ruská federace* je širokým dotazem. Otázka totiž obsahuje 4 klíčové výrazy – „předseda“, „dolní komora“, „parlament“ a „Ruská federace“. Dotaz obsahuje jen jediný klíčový výraz „Ruská federace“. Široké dotazy mají v praxi velkou nevýhodu. Uživatel musí buď dlouho procházet nalezenými výsledky, nebo nakonec upřesnit svůj dotaz, aby se dobral relevantnějších výsledků.

Oproti tomu užívání užších a tedy přesnějších výrazů je žádoucí. Pomocí nich lze totiž nalézt relevantní dokumenty ihned. V tomto ohledu je **dotaz považován za přesný**, pokud se v něm vyskytují všechny klíčové výrazy z vyjádření „úlohy“, jejíž odpověď hledáme [17]. Např. hledáme odpověď na otázku „Jaký je rok a přesné datum napadení Sovětského svazu nacistickým Německem?“. Klíčovými výrazy jsou „rok“, „datum“, „napadení“ (popř. „útok“), „Sovětský svaz“, „Německo“. Všechny tyto výrazy tedy musí být v dotazu, aby byl považován za přesný. Pokud bychom například vynechali pojem „Sovětský svaz“, vyhledávač by údaj nemusel najít. Mohl by třeba najít datum útoku nacistického Německa na Polsko či Francii. U tohoto příkladu lze také pozorovat závislost získaných informací na znalostech uživatelů o vyhledávaném tématu. Pokud by vyhledávač zadal dotaz ve formě *operace barbarossa* (což byl krycí název onoho útoku, který pod tímto označením také vešel do dějin), pak by dostal výsledek velmi rychle a přesně.

Pokud má uživatel alespoň základy některých vyhledávacích operátorů, je to pro něj jistě výhodné a může jich velmi efektivně využít. Musí si však vždy dát pozor na jejich správné použití. Chápání těchto operátorů totiž může být ve dvou vyhledávacích strojích poněkud odlišné. To, co jeden vyhledávač doplní automaticky, nemusí automaticky doplnit vyhledávač jiný. Např. operátor AND doplňuje většina vyhledávacích služeb automaticky, a tudíž ho tazatel nemusí psát. Dotaz *počítač notebook* je tedy ve většině vyhledávačů ekvivalentní zápisu *počítač AND notebook*. Takto automaticky doplněný operátor se označuje jako tzv. default operátor. Další zásadou je, aby byly operátory vždy psány velkými písmeny. Jinak je vyhledávač pochopí jako stopslova. Je na uživateli, aby svůj oblíbený vyhledávač prozkoumal, popř. zjistil funkce vyhledávače jiného. Ve všech vyhledávacích se vyskytují odkazy typu „Advanced Options“, „Search Tips“, „Search Hints“, které jsou k tomuto účelu určené.

Další příčiny nerelevantních informací souvisí již s vyhledávacími stroji. Za hlavní se považují nedokonalé vyjádření obsahu dokumentu (popř. webové stránky) klíčovými slovy a nekvalitně prováděné indexování ze strany vyhledávacího stroje. Vše osvětlím na příkladu. V nějaké databázi (např. elektronickém archivu článků jistého hudebního časopisu) budu hledat dokumenty, související s vystoupením kapel Jefferson Airplane či Ten Years After na rockovém festivalu Woodstock. Položím dotaz v podobě (*jefferson airplane OR ten years after*) *AND woodstock*. Budou tedy vyhledány články s klíčovými slovy „Woodstock“ a alespoň jedním z termínů „Jefferson Airplane“ či „Ten Years After“. V databázi může existovat článek, který se zabývá koncerty kapely Jefferson Airplane, mimo jiné i jejich vystoupením na festivalu Woodstock. Pro nás je tedy takový článek relevantní. Ale vystoupení na tomto festivalu může být z hlediska daného článku okrajové. Mohou se v něm vyskytovat popisy dalších desítek jejich koncertů. Pojem „Woodstock“ tak nemusí figurovat v klíčových slovech tohoto článku. Článek

tedy nebude nalezen, přestože je pro nás relevantní. Obdobně můžeme nalézt článek, který obsahuje požadovaná klíčová slova, avšak nebude pro nás relevantní. V tom spočívá ona nedokonalost vyjadřovací schopnosti klíčových slov [18]. Příčiny nesprávného indexování tkví v tom, že články na požadované téma mohou být indexovány velmi obecným termínem. Proto je vyhledávač, indexující pomocí klíčových slov, nenajde [19].

3.1.4 Hledáme relevantnější informace, aneb jak správně vyhledávat?

Některé z výše zmíněných důvodů nerelevantních informací se dají snadno ošetřit. Obecně lze říci, že relevantnější výsledky dostaneme, pokud budeme vyhledávat opačným způsobem, než jsem nastínil v předešlé podkapitole. Ale tak jednoduché to bohužel není.

Uživatelé nepohlíží dále než na 3. stránku s výsledky. To je jedno z klíčových zjištění. Jenže uživatel si nikdy nemůžeme být 100 % jist svým dotazem. Fakt, že nás první 3 stránky neuspokojí, nemusí být nutně důkazem toho, že se hledaná informace v databázi nevyskytuje. Může to být rovněž důsledek špatně položeného dotazu. První rada tedy je, aby uživatelé pohlédli dále než na 3. stránku s výsledky. Vždyť není nutné, aby všechny odkazy také procházeli.

Vyhledávací systémy se vzájemně liší, a tak i podoba správného dotazu může být v každém z nich poněkud odlišná. Současný trend spěje ke sbližování všech systémů a pro strukturu dotazů to platí dvojnásob. Uvedu tak nyní pouze obecné zásady při tvorbě dotazu. Konkrétní rady k 5 srovnávaným vyhledávačům zmíním v kapitole věnované jim samotným.

Základem úspěchu vyhledání relevantních informací je vhodná volba klíčových slov v dotazu. Právě této fázi je třeba věnovat největší pozornost. Neuvážená klíčová slova často vedou k velkému množství irelevantních hitů [20]. Obvyklou chybou dotazů je rovněž nedostatečný počet klíčových slov. V minulé podkapitole bylo řečeno, že průměrný počet klíčových slov na jeden dotaz jsou pouhá 2 slova. Pokud však hledáme něco konkrétního, 2 slova nemohou stačit. Doporučená horní hranice je 8 klíčových slov. Další zásadou je, abychom vynechali tzv. **stopslova**. Jedná se o v běžném jazyku velmi často užívaná slova, která však do rešeršního procesu nepatří. Nenesou totiž žádnou či téměř žádnou informaci, proto jsou vyhledávacími stroji ignorována (pokud však nejsou součástí dotazu považovaného za frázi). Proces vyhledávání je tak rychlejší. Stopslova jsou například spojky, předložky, přivlastňovací zájmena apod. Pro orientaci uvedu několik příkladů českých stopslov: a, mezi, pod, dále, můj, jí, jsem atd. Anglickými stopslovy jsou např. I, a, at, be, the, this, how, in aj. Pro úspěšné vyhledávání je také vhodné vědět, jak je v našem vyhledávací chápána interpunkce a zdali se rozlišují velká a malá písmena. Většina vyhledávačů velká a malá písmena nijak nerozlišuje. Výjimku tvoří booleovské a proximitní operátory. Ekvivalentní jsou tedy dotazy *počítač*, *poČíTač* či *POČÍTAČ*. Obecně se doporučuje psát vše malými písmeny.

Dalším pravidlem je nepřekrývat termíny, což do značné míry souvisí s vhodným počtem klíčových slov. Pokud zadám dotaz *pavel nedvěd juventus turín*, je zbytečné doplňovat ho dalšími termíny jako „fotbal“, „fotbalový klub“ apod.

Základem úspěšného vyhledávání jsou podstatná jména. Právě ta slouží v běžné mluvě k jednoznačnému pojmenování hmotných předmětů, abstraktních pojmů a myšlenek a dalších objektů. Slovesa, přídavná jména a příslovce jsou snadno zaměnitelná s jejich synonymy, a proto je jejich použití v podobě klíčových slov nevhodné. Výjimku tvoří slova, která slouží k jednoznačné identifikaci vyhledávaného objektu, např. přídavné jméno „světová“ v dotaze 2. *světová válka*.

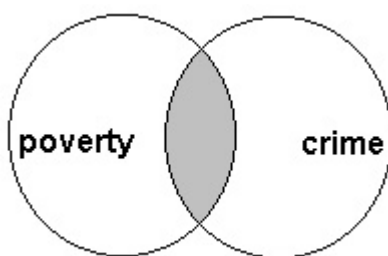
Při zadávání dotazu jsou někdy užitečné tzv. **zástupné znaky**. Nejužívanějšími zástupnými znaky jsou symboly „*“ (hvězdička) a „?“ (otazník). Zástupný znak „*“ nahrazuje libovolný počet (tj. 0 a více) jakýchkoli znaků. Používá se převážně ke zkracování slov. Požadujeme-li vysokou úplnost vyhledávání (tedy co nejvíce informací o hledaném tématu), je pro nás tento znak ideální. Budeme například hledat informace o pojmu „právo“. Pokud zadáme dotaz ve formě *práv**, pak nalezené výsledky budou obsahovat pojmy „právo“, „právní“, „právně“, „právnícký“, „práva“

apod. To jsou pojmy, které nás bezesporu budou zajímat, poněvadž patří do oblasti hledaného zájmu. Zástupný znak nám zde pomohl ve vyhledání relevantních termínů, které se vyskytují v jiném gramatickém tvaru, nebo mají s našim termínem stejný kořen. To je ostatně největší výhodou zástupného znaku „*“. Spolu s těmito pojmy však budou nalezeny dokumenty, obsahující výrazy jako „pravidlo“, „pravice“, „pravoslaví“, „pravoslavný“ (některé vyhledávače totiž interpunkci nerozlišují). Z příkladu tedy plyne velká nevýhoda zástupného znaku „*“ – kromě očekávaných rozšíření kořene „práv“ jsme dostali množství pojmů, které s hledaným tématem absolutně nesouvisí. Znak „*“ můžeme umístit také doprostřed slova. Příklad ukázal, že použití „*“ není vždy prospěšné. Jeho užití je tedy třeba vždy pečlivě zvážit. Obecně je vhodné použít tohoto symbolu u termínů s delším kořenem. Možnost vyhledání nesouvisejících slov se tím zmenší. Vyhledávací služby, podporující tento symbol, vyžadují určitý minimální počet znaků (obvykle 3) před jeho výskytem. Já osobně nemám s použitím tohoto symbolu dobré zkušenosti, a tudíž bych ho nedoporučil. U některých vyhledávačů totiž nereprezentuje tento zástupný znak nic víc než jediné slovo. Pokud ho tedy uvedu, nedostanu o nic víc výsledků, než kdybych použil jen samotné slovo. Takto chápe symbol „*“ také nejrozšířenější vyhledávač Google. Podobně je na tom Windows Live Search. Vyhledávač Exalead tento symbol naopak využívá. Z českých vyhledávačů symbol podporuje např. Morfeo.

O hodně užitečnější se jeví zástupný symbol „?“ , který slouží jako substitute žádného nebo právě jednoho znaku [21]. Symbol je velice výhodný například tehdy, pokud neznáme přesné hláskování nějakého pojmu, popř. jména. Víme třeba, že se námi hledaná osoba jmenuje Carlson či Carlsen. Pak jednoduše zadáme dotaz ve formě *carls?n*. Symbol „?“ nemůže stát na prvním místě klíčového slova. Výhodou je, že dva symboly „?“ reprezentují 2 znaky, 3 symboly „?“ reprezentují 3 znaky atd. Bohužel ani tento zástupný symbol není často vyhledávači podporován.

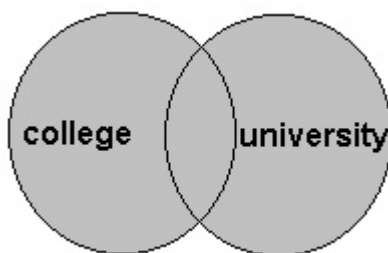
Na tomto místě bych rád zmínil užitečnost operátorů pro účinné a relevantní vyhledávání. Uživatelé by měli znát alespoň ty základní, mezi něž patří především booleovské operátory. **Booleovské operátory** definují vztah mezi slovy či skupinami slov [22]. Jsou určeny především ke zvýšení přesnosti vyhledávání. Existují 3 základní booleovské operátory: AND, OR a NOT. Při kombinování booleovských frází se dále používají kulaté závorky () ke znázornění pořadí, v jakém se mají uvažovat vztahy vyjádřené zmíněnými spojkami. Použití závorek je intuitivní. Lze uplatnit stejné myšlení jako při znázorňování pořadí výrazů v matematických rovnicích.

Spojka AND je určena k zúžení vyhledávání. Uživatel pomocí ní sděluje, že chce vyhledat záznamy obsahující všechna slova, která s ní oddělil. S pomocí množinových diagramů si ji lze představit jako na obrázku 6 na následující straně [23]. Obrázek odpovídá informačnímu požadavku najít vztah mezi chudobou (poverty) a zločinností (crime). Dotaz formulujeme jako *poverty AND crime*. Při konstrukci dotazu nezáleží na pořadí zapsaných slov. Výsledek bude vždy stejný. Zde tedy hledáme dokumenty, které obsahují obě slova zároveň. Šedě vyšrafovaná plocha odpovídá dokumentům, které tuto podmínku splňují. Ve výsledcích vyhledávání tedy nebudou žádné záznamy, kde se vyskytuje pouze výraz „poverty“ či pouze výraz „crime“. Uvědomme si, že spojka AND neříká nic o umístění či vzájemné provázanosti daných výrazů. Poukazuje pouze na jejich současný výskyt. Tento problém řeší až operátor NEAR. Počet hitů, nalezených na dotaz pomocí spojky AND, bude samozřejmě menší, než kdybychom zadali termíny jednotlivě. Čím vícekrát použijeme spojky AND (2 spojky AND mezi 3 slovy apod.), tím méně hitů logicky dostaneme. V některých vyhledávačích lze použití spojky AND zjednodušit. K tomu slouží symbol „+“ (plus), který se vkládá před slovo, jehož výskyt požadujeme. Zápis dotazu *+poverty +crime* je tedy ekvivalentní zápisu *poverty AND crime*.



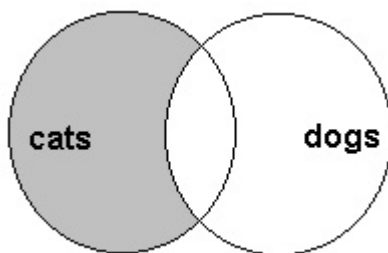
obr. 6 – Grafické znázornění spojky AND

Spojka OR je naopak určena k rozšíření vyhledávání. Sděluje vyhledávači, že si uživatel přeje získat záznamy, obsahující jakékoli ze slov, které spojkou oddělil. Místo spojky OR lze též použít operátor „|“ (svislice), který je s ní ekvivalentní. Význam spojky OR znázorňuje obrázek 7 [24]. Úlohou je nalézt informace o vysoké škole (college). Dotaz však položíme ve formě *college OR university*. Slovo „university“ je totiž synonymem ke slovu „college“. Pořadí slov je opět lhostejné. Takto položeným dotazem získáme dokumenty, kde se vyskytuje alespoň jedno z těchto klíčových slov. Relevantní pro nás budou dokumenty obsahující buď termín „college“, nebo termín „university“ (nebo oba najednou). Této podmínce odpovídá šedě vyšrafovaná plocha. Dalším synonymem těchto 2 slov je slovo „campus“ (v americké angličtině taktéž možné označení vysoké školy). Mohli bychom tedy položit dotaz *college OR university OR campus*. Získali bychom tak ještě větší množství relevantních dokumentů. Pokud tedy vyhledáváme určitý pojem, vezměme v úvahu všechny jeho synonyma a dotaz položíme ve formě těchto synonym, oddělených operátorem OR. To samé provedeme, pokud by k hledanému termínu existovaly nějaké alternativní tvary. Například ve starších knihách a publikacích se slovo „univerzita“ píše jako „universita“. Čím vícekrát užijeme spojku OR, tím více výsledků dostaneme.



obr. 7 – Grafické znázornění spojky OR

Operátor NOT je stejně jako AND určen k zúžení vyhledávání. Vyjadřujeme jím požadavek, aby vyhledávač našel dokumenty, které neobsahují slova za termínem NOT (resp. slova „znegovaná“ termínem NOT). Význam této spojky ilustruje obrázek 8 na následující straně [25]. Úkol zní nalézt informace o kočkách (cats), ale vyvarovat se jakýchkoli informací o psech (dogs). Dotaz položíme formou *cats NOT dogs*. Chceme tedy nalézt dokumenty obsahující pouze jeden z těchto termínů (zde cats). Kritéria splňují šedě vyšrafované dokumenty. V nich se neobjeví žádné dokumenty, ve kterých se vyskytuje slovo „dogs“, byť by obsahovaly termín „cats“. Podobně jako u AND funguje v některých vyhledávačích i u spojky NOT možnost zkráceného zápisu, tentokrát pomocí symbolu „-“ (minus). Dotazy *+cats -dogs* a *cats NOT dogs* jsou tedy rovnocenné.



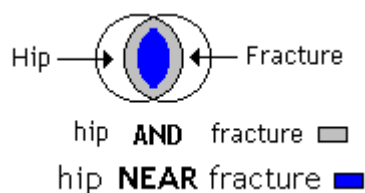
obr. 8 – Grafické znázornění spojky NOT

Ze zmíněných charakteristik plynou následující doporučení. Pokud na zadaný dotaz dostaneme příliš mnoho výsledků, je to pro nás impulsem, abychom přidali k dotazu další, upřesňující klíčové slovo a obě oddělili spojkou AND. Pokud vyhledávač vrátí příliš mnoho výsledků, obsahujících i odkazy na zcela irelevantní témata, měli bychom se pokusit vyjádřit tato témata nějakým klíčovým slovem (resp. slovy). Irelevantní odkazy pak eliminujeme operátorem NOT, který umístíme před toto klíčové slovo (resp. slova). A konečně pokud vyhledávač nalezne příliš málo záznamů, zvažme přidání dalšího klíčového slova a jeho oddělení operátorem OR. Všechny tři booleovské operátory včetně závorek podporují oba vyhledávače Exalead i Windows Live Search. Google umožňuje pouze užití operátoru OR.

Poslední pomůckou, kterou zmíním jako nástroj k účinnému vyhledávání, jsou **proximitní operátory**. Ty umožňují vyjádření požadavku na vzájemnou pozici vyhledávaných termínů [26]. Předávají vyhledávačům požadavek uživatele, aby vzdálenost mezi jeho klíčovými slovy nebyla větší než jistá mez. Mezi může být jedna fráze, jedna věta, jeden odstavec apod. Významově stojí proximitní operátory velice blízko spojce AND, poněvadž se ve výsledcích musí vyskytovat všechny klíčové výrazy.

Nejužívanějším nástrojem, radícím se mezi proximitní operátory, je bezesporu využití frází. Fráze jsou tvořeny kombinací slov, které musí být přítomny v dokumentu ve stejném pořadí, v jakém je zadal uživatel [27]. V dotazu se k jejich vyjádření využívá dvojitých („“) či jednoduchých (‘ ’) uvozovek. Jejich využití je vcelku logické. Používají se v případě, kdy vyhledáváme určité ustálené spojení, např. „2. světová válka“, „železná opona“, „studená válka“ atd. Uvnitř fráze nejsou přípustné zástupné znaky. Booleovské operátory však ano. Ve frázích je správná diakritika a absence překlepů důležitější než kdekoliv jinde. Rovněž mezera mezi slovy fráze musí být vždy jednoduchá (nikoli dvojitá). Fráze je jediným proximitním operátorem, který je podporován drtivou většinou vyhledávačů. A nejen to. Většina z nich rozpozná frázi i tehdy, pokud vynecháme uvozovky. Pokud tedy použijeme dotazu *2. světová válka* bez uvozovek, výsledky s tímto výrazem jako frází budou mezi prvními nalezenými.

Druhým nejfrekventovanějším (a také jedním z posledních používaných) proximitním operátorem je operátor NEAR (u některých vyhledávačů je jeho ekvivalentem operátor ADJ). Lze si ho představit jako určité zúžení spojky AND. Do značné míry řeší neprovázanost klíčových slov ve vyhledaných dokumentech. Vyhledávací systémy však operátor NEAR, zapsaný takto osamoceně, chápou různě. Většina z nich bere vzdálenost vyjádřenou operátorem NEAR jako 10 slov. To znamená, že klíčová slova ve vyhledaných dokumentech od sebe nesmí být vzdálena více než 10 slov. Je přitom lhostejno, v jakém pořadí se v textu vyskytují. Vztah operátorů NEAR a AND je znázorněn na obrázku 9 na následující straně [28]. Schématu odpovídá následující informační požadavek: „Jaké jsou druhy zlomenin boku lidského těla?“. Bok těla se v angličtině řekne „hip“ a zlomenina „fracture“. Uživatel tedy položil dotaz formou *hip NEAR fracture*. Modrá barva značí hity, nalezené na daný dotaz. Sjednocení šedé a modré plochy představuje výsledky, které by byly nalezeny na dotaz *hip AND fracture*. Je zřejmé, že obsah modré plochy je menší než sjednocení obou ploch. Zde je patrné zúžení operátoru NEAR oproti AND.



obr. 9 - Vztah operátorů AND a NEAR

O kolik však bude modrá plocha menší než celková? Jak jsem již podotkl, záleží to na použitém vyhledávači. Je však velice uspokojivé, že rovněž uživatel může do velikosti této plochy výrazně promluvit. K tomu mu slouží možnost upřesnění operátoru NEAR, které lze vyjádřit několika způsoby, opět závislým na zvolené vyhledávací službě. Ve většině případů jsou přípustná vyjádření NEAR/# či NEAR# (někdy se též používá operátoru WITHIN#). Symbol „#“ značí maximální přípustnou vzdálenost (ve slovech) mezi klíčovými termíny. Tak například NEAR1 znamená, že klíčová slova musí stát vedle sebe. V našem příkladě to říká následující – slovo „hip“ je vzdáleno 1 slovo od slova „fracture“. Rozdíl mezi dotazy „*hip fracture*“ (jako fráze) a *hip NEAR1 fracture* je v tom, že operátor NEAR1 připouští i opačné pořadí slov v dokumentu, kdežto fráze nikoli. Operátor NEAR2 povoluje jedno jediné slovo mezi klíčovými termíny. Do výsledků by tedy v našem příkladě byl zařazen i dokument s výskytem slov „fracture of hip“, který by pro nás byl jistě relevantní. Obdobné je to u NEAR3 a výše. Ještě bych připomenul, že pokud chci použít jakéhokoli operátoru jako slova a ne jako operátoru, musím ho dát do uvozovek. Podpora jiných proximitních operátorů než frází je značně omezená. Exalead podporuje operátor NEAR. Google ani Windows Live Search nepodporují kromě frází žádné jiné proximitní operátory.

V této podkapitole jsem zmínil všechny základní prostředky k vyhledávání. Opravdu efektivního vyhledávání však dosáhneme až jejich šikovnou kombinací. Pokud je tedy shrnu, pak základní **pravidla pro tvorbu dotazů při vyhledávání v Internetu jsou:**

- vhodně volit klíčová slova – jaká a kolik jich užít
- vynechat stopslova
- není nutno rozlišovat velká a malá písmena
- základem dotazu jsou podstatná jména
- pro úplnost vyhledávání užít zástupného symbolu „*“
- pro zpřesnění vyhledávání použít booleovských operátorů AND, OR, NOT, popř. zkrácení operátoru AND symbolem „+“ a NOT symbolem „-“
- uplatnit logické myšlení pro použití závorek ()
- při hledání fráze užít proximitního operátoru („“)
- pro nalezení provázanosti klíčových slov užít proximitního operátoru NEAR

Výše zmíněné zásady lze uplatnit jen tehdy, pokud je podporuje naše vyhledávací služba. Dotazy konstruované dle těchto zásad se již dají označit za pokročilé dotazy. Jejich osvojení je čistě záležitostí praxe a zkušeností. Na druhou stranu není vhodné konstruovat dotazy příliš složitě. Čím složitější dotaz, tím pravděpodobnější je výskyt chyby, takže i zde opatrně. Příklad pokročilejšího dotazu by mohl vypadat následovně. Vymyslel jsem následující informační požadavek: „Nalézt informace o roztržkách během studené války v 60. letech mezi Spojenými státy a Sovětským svazem, kde však nefigurovalo Německo.“ Dotaz bychom mohli položit takto:

((krize NEAR/10 „studená válka“) AND „60. léta“ AND (usa OR spojené státy americké) AND (sssr OR sovětský svaz)) NOT (německo OR nsr OR ndr)

K ozkoušení jsem použil vyhledávače Exalead. Ten jediný totiž podporuje všechny operátory, vyskytující se v daném dotaze. Příliš jsem od dotazu neočekával, přeci jen jsem ho

považoval za příliš dlouhý a kostrbatý. Bral jsem ho spíše jako vzorovou ukázkou jednoduchého vyjádření složitějšího dotazu. Navíc jsem byl skeptický k faktu, že by existoval dokument o krizích studené války, kde by v nějakém místě nefigurovalo Německo. Rovněž požadavek vzdálenosti slova „krize“ a pojmu „studená válka“ mi přišel nereálný. Výsledek mne však velmi příjemně překvapil. Hned první nalezený hit obsahuje téměř 100 % relevantní informace.

3.1.5 Vyjádření speciálních požadavků na relevantní informace, aneb pokročilé možnosti vyhledávání

V předchozí podkapitole jsem nastínil možnosti při pokládání dotazů a vyhledávání obecně. Zejména náročnější uživatelé se však s těmito nástroji nespokojí. Široké možnosti uplatnění má použití **filtrů**. Ty zvyšují především přesnost vyhledávání. Filtry jsou nezávislé na dotazu [29]. To znamená, že nějakým způsobem zužují prostor dokumentů, který bychom získali položením dotazu bez použití filtrů. Pokud je nepoužijeme, pak vyhledávač pátrá po dotaze ve všech oblastech (z časového i věcného hlediska). Praktická aplikace může být dvojitá. První možností je využití „pomocníků“ při filtrování. Ve většině vyhledávačů jsou to odkazy typu „Advanced Search“, „Pokročilé hledání“ apod. Druhou možností je zápis filtru přímo jako součást dotazu, což však vyžaduje znalost syntaxe požadovaných filtrů. Pokud totiž filtr použijeme špatně, vyhledávač většinou bere dotaz jako frázi. Pokud syntaxi známe, pak obecné schéma dotazu vypadá následovně – *nazev_filtru:jeho_hodnota_další_klíčová_slova*. Mezi názvem filtru, dvojtečkou a hodnotou filtru nesmí být mezera. Hodnotou filtru většinou může být i pokročilý dotaz či fráze. Některé filtry (a je jich poměrně dost) jsou dostupné pouze zápisem v dotazu, tj. v možnostech pokročilého vyhledávání je nenajdete. Název operátoru se navíc může lišit v závislosti na vyhledávací službě.

Zmíním alespoň ty nejpoužívanější filtry, kterými jsou „intitle“, „inurl“, „link“, „site“ a „filetype“. Drtivá většina vyhledávacích služeb tyto filtry podporuje. A co víc, u většiny zahraničních vyhledávačů mají tyto operátory shodné názvy. Filtr „intitle“ vyhledává záznamy, které obsahují požadovaná klíčová slova v titulku stránky, což je obvykle text v záhlaví okna prohlížeče. Pro znalé HTML kódu to znamená, že klíčové slovo se musí nacházet v elementu „title“. To však neplatí u automaticky generovaných stránek, které takový titulek mít nemusí. Např. dotaz ve formě *intitle:válka_adolfa_hitler* by měl najít dokumenty s výrazem „válka“ v titulku a jménem „Adolf Hitler“ kdekoli ve stránce.

Filtr „inurl“ vyhledává klíčová slova v URL adresách [30], tj. v názvu subdomény (např. vse), domény nejvyšší úrovně (např. .cz) nebo v umístění daného dokumentu (např. /obecne/o_skole.php). Dotaz *inurl:vse.cz_kasik* tedy vyhledá všechny dokumenty obsahující „vse.cz“ v URL adrese a mé příjmení kdekoli v textu. Tento filtr je vhodný zejména při hledání důvěryhodných zdrojů (viz další podkapitola). Filtr „link“ vyhledává stránky s hypertextovými odkazy na požadované umístění. Dotaz *link:www.vse.cz* tedy vyhledá všechny stránky, obsahující odkazy na oficiální stránky naší školy. Mezi výsledky budou samozřejmě i dokumenty, nacházející se na samotných stránkách VŠE. Filtr je výhodný například tehdy, chce-li si majitel své domény vyhledat všechny stránky, které se na ni odkazují. Z citační analýzy totiž víme, že odkazy, na které jsou často uváděny reference, jsou většinou kvalitní. Také tento filtr tedy lze použít k ověření důvěryhodnosti získaných informací.

Velmi výhodným filtrem je filtr „site“. Ten hledá klíčová slova na stránkách, které hostují na konkrétním serveru, nebo náležejí k určité doméně. Dotaz *site:www.nhl.com_jagr* vyhledá všechny výskyty slova „jagr“ na oficiálních stránkách NHL. Filtr je výhodný zejména tehdy, pokud nemá daná webová stránka vlastní vyhledávací systém. Některé stránky naopak využívají služeb vyhledávačů jako svého vlastního vyhledávacího systému. Tak je tomu i u oficiálních stránek naší školy, jejichž vyhledávací systém používá služeb Googlu. Posledním filtrem, který zmíním, je „filetype“. Jak již název napovídá, jedná se o filtr, který vyhledává klíčová slova v souborech s požadovanou příponou. Dotaz *filetype:pdf_vse OR vysoká_skola_ekonomická OR*

vysoká škola ekonomická v praze tedy nalezne všechny výskyty názvu naší školy v souborech s příponou .pdf.

3.2 Důvěryhodnost

3.2.1 Co je důvěryhodnost a co důvěryhodnost webu?

Důvěryhodností obecně je označována uvěřitelnost sdělení či zdroje a také tendence nezávislého pozorovatele uvěřit danému sdělení či zdroji [31]. Většina badatelů se shoduje na dvou základních vlastnostech důvěryhodnosti. Těmi jsou pravdivost a odbornost informací [32]. **Důvěryhodnost webu** je informační spolehlivost a poučitelnost (tj. schopnost poučit) nějaké konkrétní webové stránky [33]. Informace je důvěryhodná, pokud se jí dá věřit. S důvěryhodností stránky roste počet stránek, které se na ni odkazují. Tím roste i její prestiž. Část věnovaná relevanci obsahovala doporučení, jak v Internetu najít informace, odpovídající našim požadavkům. Tato kapitola se bude zabývat problémy souvisejícími s tím, jak z nalezených informací vybrat ty důvěryhodné.

3.2.2 Je Internet důvěryhodný?

Na tuto otázku nelze jednoznačně odpovědět. Správná odpověď by měla znít, že pro některé účely ano, pro jiné naopak ne. Například obchodníci s akcemi činí na základě informací z Internetu rozhodnutí, ve kterých často jde o miliony dolarů. Pro podobné účely tedy Internet důvěryhodným je. Nicméně Internet může být též zcela nevhodným zdrojem informací. Jedná se totiž o anonymní médium. Svůj web si dnes může vytvořit prakticky kdokoli. A tak se zde vyskytují tisíce stránek s podvodným obsahem. Fakt, že neexistují žádné bariéry k umístění informací na Internet, způsobuje, že se na něm objevují chybné informace a reklamy, slibující neuskutečnitelné. To na jedné straně staví uživatele Internetu před úkol rozeznat informace správné od informací nesprávných. Na straně druhé stojí firmy či jedinci, kteří se snaží zlepšit své stránky tak, aby patřily mezi ty důvěryhodné. Jejich snahou totiž je, aby se uživatelé na tyto stránky vraceli. A nejen to. Často se snaží uživatele pobídnout, aby se na jejich stránkách zaregistrovali, nakoupili, klikli na reklamy sponzorů, nebo aby si stránku uložili do záložek. Důvodů je skutečně bezpočet. A právě důvěryhodnost poskytnutých informací jim v dosažení těchto cílů pomůže. Z předchozích vět je zřejmé, že důvěryhodnost webu je klíčová hlavně pro internetové obchody. Posiluje totiž jejich jméno a značku.

3.2.3 Jak ohodnotit důvěryhodnost webu?

Předpokládáme, že jsme pomocí technik z minulé kapitoly našli určité informace. Jak z nich ale vybrat informace zaručeně správné a tudíž důvěryhodné? Na to se teď pokusím odpovědět. Jde především o to, aby byl důvěryhodný samotný zdroj, ze kterého informace pochází. Jinak řečeno, důvěryhodná musí být ona webová stránka. Problémem je, že faktory ovlivňující důvěryhodnost webu jsou velmi často subjektivní záležitosti. Naprosto přesné a jednotné metody zjišťování důvěryhodnosti tak neexistují, tudíž i v této oblasti hrají prim průzkumy chování uživatelů. Nejdůležitější je zjistit, co vše berou uživatelé v úvahu, pokud se rozhodují, čemu uvěřit a čemu naopak ne. Dalším problémem je ověřování informací. Pokud něco uživatelé najdou, mají snahu si to ověřit z jiného zdroje? Podobnými otázkami se již od roku 1998 zabývá tým pod vedením BJ Fogga ze Stanfordské univerzity v Kalifornii. Na téma důvěryhodnosti jeho tým již uskutečnil několik studií. Ta poslední je sice stará 4 roky, avšak změna chování průměrného uživatele je záležitostí dlouhodobou. Z toho důvodu budu čerpat právě z ní [34]. Potěšitelné jistě je, že počítačové zkušenosti hrají v procesu rozhodování o důvěryhodnosti podstatně menší roli než v procesu vyhledávání relevantních informací.

Zmíněná studie provedla výzkum na 2 684 respondentech, kteří hodnotili důvěryhodnost webových stránek z celkem desíti oblastí. Z každé z desíti oblastí bylo vybráno 10 webových stránek, tj. celkem se ve studii pracovalo se 100 weby. Každý respondent hodnotil dvě stránky. Drtivá většina z nich byli občané Spojených států s průměrným věkem 39,9 let. Badatelé se snažili zjistit, jaké rysy webových stránek uživatelé berou v potaz, pokud hodnotí jejich důvěryhodnost. Deset nejdůležitějších rysů nyní uvedu. Čísla indikují, kolik procent uživatelů zmínilo ten který rys.

Na 1. místě se umístil **design a vzhled stránky** se 46,1 %. Tento pojem zahrnuje především vizuální vzhled, včetně grafického rozvržení, typů písma, využití celkové plochy obrazovky, množství bílých míst, obrázků, barevných schémat apod. Hlavní myšlenkou je, že stránka, která vypadá „profesionálně“, je u uživatelů důvěryhodnější. Přitom ani tak nezáleží na jejím věcném obsahu. Grafický vzhled stránky je to první, čeho si na stránce všimneme. Buď nás doslova uhrane skvělým vyvážením barev a mistrným využitím barevných schémat, nebo nás naopak udeří do očí její nedbalý a někdy až infantilní design. Na toto téma jsem našel jednu větu, která trefně vystihuje fakt, proč je na 1. místě hodnocen právě vzhled stránky [35]: „Uživatelé mají v podvědomí, že pokud něco dobře vypadá, pak to také dobré je. Tudíž je to i důvěryhodné.“ Důvodem je, že návštěvník stráví na stránkách v průměru velice krátkou dobu. Za tak krátký úsek je právě design webu to jediné, co mu utkví v paměti.

Na 2. místo postavili uživatelé **strukturu informací** (28,5 %), tedy samotnou věcnou náplň stránky. Do této oblasti patří zejména organizace a ucelenost informací a jejich provázanost. S tím souvisí i možnost navigace. Tedy jak těžké je dostat se pomocí odkazů k informacím, které nás zajímají. Stránky se snadnější navigací označili uživatelé jako více důvěryhodné. Naopak stránky, které se snaží mít všechny informace hned na úvodní stránce, nepůsobí příliš důvěryhodně.

Jako 3. nejdůležitější faktor důvěryhodnosti bylo vyzorováno **informační zaměření stránky** (25,1 %). Stránky s jasným informačním ohniskem jsou důvěryhodnější než weby bez jasného cíle zaměření. Na nich se často vyskytuje množství nadbytečných informací, které uživatele od těchto stránek odtahují. Stránky s jasně formulovaným informačním posláním jsou většinou určeny skupině uživatelů s užším cílovým zaměřením. Například oficiální stránky nějakého fotbalového klubu jsou určeny speciálně fanouškům tohoto klubu a jejich jasným zaměřením je poskytovat informace o tomto klubu. Tyto stránky tedy budou u uživatelů-fanoušků důvěryhodné. Pravým opakem mohou být weby s nabídkami rychlého zbohatnutí či bezbolestného shození přebytných kil. Co však mají takové stránky uživateli sdělit, zůstává záhadou.

Celých 15,5 % uživatelů zmínilo jako podstatný faktor důvěryhodnosti **základní motiv stránky**. Bylo zjištěno, že stránky ztrácí věrohodnost, pokud je jejich hlavním motivem prodej a příjem z tohoto prodeje. Naopak za důvěryhodné byly považovány stránky s motivy, které byly ze strany uživatelů označeny za „hodné obdivu“. Pod tímto pojmem se má na mysli úsilí webu, aby uživatelé našli skutečně to, co hledají. Důvěryhodné stránky mají sloužit hlavně uživatelům. Opakem jsou stránky typu „dejte nám vaše peníze a běžte“, kde firma nejdříve vychválí svůj produkt a někde poblíž je ihned k dispozici odkaz k jeho objednávkě.

Také **užitečnost informací** je pro uživatele znakem důvěryhodnosti (14,8 %). S tím souvisí především zajímavost informací pro uživatele. Dalším aspektem je **přesnost informací** (14,3 %). Každý uživatel má jisté vědomosti. Pokud najde informaci, která odporuje jeho poznatkům, vzbudí v něm pochyby. A to nejen vůči informaci samotné, ale také vůči důvěryhodnosti stránky. Příkladem mohou být stránky s osobními (a tudíž subjektivními) názory či vědecky nepodloženými domněnkami. **Známé jméno a reputace stránky** (14,1 %) jsou pro uživatele také důležité. Neznámé stránky jsou nedůvěryhodné. Neznámým jménem se myslí především část URL adresy, resp. doména. Opakem jsou weby, o kterých uživatelé již někdy slyšeli, nebo se s nimi setkali. Příkladem mohou být světoznámá jména (např. Coca-Cola, CNN, Microsoft aj.). Uživatelé věří, že by tyto společnosti nikdy nedovolily, aby bylo jejich jméno podepsáno pod něčím pochybným. Znakem nedůvěryhodnosti jsou **reklamy** (13,8 %). Každý uživatel ví, že bez reklam by takřka žádná firma neobstála. Proto také většina uživatelů nezavrhne reklamy na stránkách úplně.

Každému jsou sice nepříjemné, ale považují to za jakési nutné zlo, bez něhož by nabízených služeb nemohli využívat. Nic se však nesmí přehánět. Nástroje jako například pop-up okna snižují důvěryhodnost stránek velmi strmě. **Informační zaujatost** (11,6 %) je dalším rysem, snižujícím důvěryhodnost zdroje. Posledním z deseti hlavních rysů je **tón psaní** (9 %), vyskytujícího se na stránce. Jedná se zejména o použití hrubších a slangových slov, které činí stránku nedůvěryhodnou. Pominu-li slova nejnižší mravní kategorie, pak se může jednat o slova typu „rozcupovat“, „bastard“, „idiot“, „poldové“ aj. Tato slova ani nemusí být užita nijak zákeřně. Často se však užívají k dodání senzacechtivosti. Podobné zdroje nicméně připomínají bulvár, a tak jsou nedůvěryhodné.

Pokud tedy uživatel hodnotí důvěryhodnost webové stránky, zmíněných 10 rysů je pro něj nejpodstatnějších. Pro názornost jsem je sestavil do tabulky 3 [36], včetně několika rysů dalších.

Zdroj (ne)důvěryhodnosti stránky	Výskyt (v %)
Design a vzhled	46,1
Struktura informací	28,5
Informační zaměření	25,1
Základní motiv	15,5
Užitečnost informací	14,8
Přesnost informací	14,3
Znamé jméno a reputace	14,1
Reklamy	13,8
Informační zaujatost	11,6
Tón psaní	9,0
Původ sponzora	8,8
Funkčnost	8,6
Služby zákazníkům	6,4
Předěšlá zkušenost se stránkou	4,6
Srozumitelnost informací	3,7

tab. 3 - Posouzení důvěryhodnosti webových stránek

Pokud tedy provedeme zhrubý pohled na stránku, pak **základními pomůckami pro posouzení její důvěryhodnosti** jsou:

- profesionální vzhled
- srozumitelná navigace
- na 1. pohled je patrné, o jaký typ stránek se jedná (firemní x individuální osoby)
- zajímavé a pravdivé informace
- URL adresa stránky je povědomá
- nevyskytují se reklamy a pop-upy
- absence chyb všeho druhu – typografických, pravopisných, faktografických, v URL adrese
- čerstvé informace – aktuálnost údajů, žádné mrtvé odkazy
- objektivní informace – hlavně u recenzí a informací o produktech
- mírný a přátelský tón sdělení

Vlastnosti, které uživatelé požadují po důvěryhodných stránkách, jsou rovněž závislé na povaze a zaměření webů. Například u firem z oblasti internetového nakupování požadují uživatelé za rozhodující reputaci a dobré jméno firmy. U stránek poskytujících zpravodajské služby dávají přednost nezaujatosti informací. U stránek cestovních kanceláří požadují kvalitní služby

zákazníkům. U vyhledávacích služeb byly nejčastěji zmíněny faktory struktury informací, funkčnosti a množství reklam. Na základě těchto poznatků lze formulovat jisté zásady při tvorbě webových stránek.

3.2.4 Jak učinit svůj web důvěryhodnějším?

Pokud něco publikujeme na Internetu, zřejmě tak činíme za nějakým účelem. Patrně chceme, aby byly naše stránky navštěvovány internetovými uživateli. A co víc, aby jim přinášely něco nového, něčím je obohacovaly. K tomu však musíme poskytnout kvalitní informace. Aby jim však návštěvníci věřili, musí považovat náš web za důvěryhodný, což musíme zajistit právě my. Pokud tedy jako tvůrci webových stránek usilujeme o důvěryhodnost webu, měli bychom se držet jistých zásad. Tyto zásady vyplývají ze studií, podobných té z minulé podkapitoly. Formují se jako reakce na vypořizované zdroje nedůvěryhodnosti. Je-li hlavním kritériem hodnocení důvěryhodnosti ze strany uživatele vzhled stránky, pak musíme našim stránkám vštěpit takový vzhled, aby vypadaly důvěryhodně. Podobné je to u dalších kritérií.

Stejný tým, který provedl studii zmíněnou v minulé podkapitole, navrhl několik doporučení pro tvorbu důvěryhodného webu [37]. Hlavní zásadou je **věnovat pečlivou přípravu a trpělivost vzhledu naší stránky**. Jak bylo řečeno, je to ta první věc, která návštěvníka na našem webu zaujme, popř. odradí. Jedná se o širokou škálu prvků. Od barevných kombinací, typů a velikostí písem počínaje, přes obrázky, loga a navigační lišty, konzistencemi a poměrem ve velikostech konče. Je také velice důležité, aby vzhled stránky odpovídal jejímu zaměření. Stránka by měla vypadat co nejprofesionálněji. Není však nutné dokazovat návštěvníkům, jak umíme využívat nejmodernějších nástrojů tvorby webu jen proto, aby stránka vypadala profesionálně. Musíme najít přirozený kompromis mezi snadným ovládáním stránky a její užitečností. Další zásadou je, abychom **poskytli návštěvníkům možnost ověření správnosti informací**, publikovaných na našem webu. Jde hlavně o to, abychom uváděli citace a reference na zdroje našich informací. Každé tvrzení musí být něčím podloženo. Pokud např. uvedeme, že se nám podařilo za poslední rok zdvojnásobit produkci zboží, pak musíme být schopni doložit to nějakým důkazem. Jak jinak pozná potenciální zákazník, že mluvíme pravdu? Velmi často si přitom uživatel nedá tu práci, aby vše ověřil. Nicméně každý odkaz na zdroj informací zvyšuje důvěryhodnost. Neodkazujeme však na stránku, která je sama nedůvěryhodná a vyvarujeme se tzv. mrtvých odkazů. Mohlo by to otřást také s naší důvěryhodností.

Pokud se jedná vlastní firmu, je vhodné **doložit její fyzickou existenci a další aktivity**. Nejlepším způsobem je uvést naši adresu, telefonní číslo, email, sídlo firmy a její stručnou historii. Důvěryhodně rovněž působí fotografie sídla firmy, popř. zaměstnanců. Takové údaje se většinou uvádí do sekce „Kontakty“ nebo „Kontaktujte nás“. Snažme se potenciální návštěvníky přimět, aby nás kontaktovali, resp. přesvědčit, že se kontaktu s nimi nebojíme. Tuto zásadu mnozí podceňují. Stránka s kontakty jistě nebude ta první, kterou si uživatelé budou chtít přečíst. Ale může být tou poslední, kde se uživatelé budou rozhodovat, zda s námi mají či nemají uzavřít obchod. S tím rovněž souvisí údaje na podporu tvrzení, že naše firma nebankrotuje. Musíme naopak prokázat, že se vyvíjíme a počet našich klientů stále roste. Důvěryhodně rovněž působí údaje o našich sponzorech a partnerech. Odkaz na mateřskou (resp. dceřinnou) společnost je samozřejmostí. Chybou rozhodně nejsou ohlasy spokojených zákazníků, popř. diskusní fóra. Další zásadou je **pravidelná aktualizace naší stránky**. Tento aspekt budu probírat v další podkapitole. Snažme se **vyvarovat reklam**. To se týká zejména pop-up oken. Mějme na paměti, že každé kliknutí navíc uživatele unavuje. Pokud tedy musíme mít reklamy, ať alespoň nikoho neotravují. Snažme se **vyvarovat typografických chyb a překlepů**, které rovněž snižují důvěryhodnost. Texty by měly být přátelské, srozumitelné a jednoznačné.

„Surfování“ se velmi často podobá monologu, který vede webová stránka s uživatelem. Je to Internet, který má 100 % kontrolu nad tímto rozhovorem. Pokud tedy uživatelé pojmou podezření, že je naše webová stránka nedůvěryhodná, pak nemáme šanci se nijak bránit. Musíme tedy zajistit,

abychom na stránku **umístili odpovědi na všechny potenciální otázky**, které by návštěvník mohl mít. Často se jedná o choulostivé záležitosti, o kterých bychom raději pomlčeli. Abychom však dostali naši důvěryhodnosti, tak zkrátka nemůžeme. Týká se to např. údajů o reklamaci zboží, zacházení s osobními údaji apod. K tomuto účelu může dobře posloužit kniha návštěv a stížností. Nesmíme samozřejmě opomenout celkovou **funkčnost webu**. Nebude-li stránka fungovat, působí to značně nedůvěryhodně. Rovněž pokud návštěvník nedostane odpověď (a nejlépe rychlou) na svůj dotaz, nebude se již stránkou dále zabývat. Na závěr uvedu jednu protřelou větu, která se týká důvěryhodnosti ve všech oborech lidské činnosti. Získání důvěry je proces dlouhodobý, její ztráta však může přijít během chvilky.

3.2.5 Důvěryhodnost a vyhledávací stroje

V současné době jen málokdo pochybuje o důvěryhodnosti samotných vyhledávacích strojů. Jejich uživatelé si však neuvědomují, že záznamy, nalezené po zadání dotazu, mohou být někdy nedůvěryhodné [38]. Mnoho vyhledávacích strojů dnes totiž od inzerentů přijímá nemalé částky výměnou za vyšší umístění ve výsledcích vyhledávání. Někdy jsou takové hity označeny jako sponzorské či partnerské. Tato reklama činí značné problémy metavyhledávačům, které integrují několik vyhledávacích služeb do jedné a z každé z nich nabízejí několik prvních výsledků. Jenže právě ony mohou být těmi sponzorskými, takže metavyhledávač někdy vrátí samé sponzorské výsledky. A co hůř, nijak neodliší, že se jedná právě o ně. Uživatel tak často důvěřuje informaci, která není zcela objektivní, tudíž ani důvěryhodná. Ovšem jsme opět u chování uživatelů. Málokdo si informaci ověří z více zdrojů. Některé metavyhledávače však po vyhledání dotazu nabídnou odkaz typu „O hledání“. Je dobré si ho prohlédnout, neboť se zde nacházejí informace o proběhlém procesu vyhledávání. Ty nám často dopomohou k rozhodnutí o důvěryhodnosti nalezené informace.

Na samý konec doplním využití filtrů „inurl“ a „link“. Jak jsem zmínil v úvodu této podkapitoly, jedna ze záruk důvěryhodného zdroje je jeho URL adresa. Pokud například hledám informace o nějakých českých stránkách věnovaných sportu, mohu zadat dotaz *inurl:sport.cz*. Vyhledávač mi najde všechny výskyty „sport.cz“ v názvu URL adresy. Je velice pravděpodobné, že pokud bude mít stránka v URL adrese tento pojem, pak bude nejen relevantní, ale i důvěryhodná. Navíc bude-li mít daný zdroj vlastní doménu, jeho důvěryhodnost se ještě zvýší. Filtrování „link“ je vhodný zejména k ověření důvěryhodnosti zdroje. Pokud se na daný web odkazuje velké množství stránek, zdroj bude s největší pravděpodobností důvěryhodný.

3.3 Aktuálnost

3.3.1 Co je aktuálnost a jak na ni pohlížet?

Aktuálností se rozumí čerstvost, novost, nedávnost. Pokud je něco aktuální, není to zastaralé, nýbrž čerstvé a nové, často stále ještě platné. Aktuálností informací v Internetu se tedy má na mysli jejich čerstvost. Čerstvost v tom smyslu, že byly nedávno vloženy či změněny. Této změně se říká aktualizace informací. Aktuálnost není z hlediska vyhledávání informací v Internetu jednoznačná. Záleží především na povaze informačního požadavku. Pokud hledáme informace o 2. světové válce, pak nepovažujeme za nutné, aby byl daný zdroj často aktualizovaný. Dokonce nám nemusí vadit, že zde informace leží již 2 roky bez jakéhokoli zásahu. Problém však nastává, pokud hledáme něco novějšího. V takových případech již můžeme hovořit o aktuálnosti informací jako problému. Co však znamená ono novější? Nedávno jsem si v nejmenovaném CD a DVD bazaru objednával jedno CD přes Internet. Nejprve jsem si ho zarezervoval a hned příští den se vydal na nákup. Jaké bylo mé překvapení, když mi prodávající sdělil, že ho tam již nemají. Odůvodněním bylo, že jejich internetový seznam CD moc často neaktualizují. Aby se však vyvaroval podobných nedorozumění, musel by bazar aktualizovat seznam každý den. Nebo ještě lépe každou hodinu. Naopak sportovnímu serveru, zabývajícemu se výsledky fotbalových zápasů, by stačilo aktualizovat

informace každý týden po skončení ligových kol. U aktuálnosti tedy záleží nejen na povaze informačního požadavku, ale i na zaměření daného informačního zdroje. Aktuálnost informací jako problém souvisí do značné míry s důvěryhodností. Pokud se na webu nebudou vyskytovat informace aktuální, web nebude důvěryhodný. Aktuálnost je tedy dalším z požadavků na důvěryhodný web. Pro srozumitelnost následujícího textu považuji za nutné stanovit dvě úmluvy. Pod pojmem aktuální informace budu rozumět informace nedávno vložené či změněné. Pod pojmem aktuální sdělení budu mít na mysli časovou platnost samotného významu informací.

3.3.2 Aktuálnost jako problém

Všechny informace v Internetu bych podle aktuálnosti rozdělil do 4 kategorií. Jednak jsou to informace, které nejsou aktuální a ani samotné sdělení není aktuální. Příkladem může být dva roky starý článek o Napoleonově tažení do Ruska. Druhou kategorií jsou informace neaktuální, avšak se stále platným sdělením, které se však může změnit. Jako příklad uvedu stránku s kontakty na nějakou firmu. Firma například sídlí již dva roky v Praze. Pokud se však přestěhuje či změní telefonní číslo, bude muset své kontakty aktualizovat. Třetí kategorii představují informace aktuální, avšak s neaktuálním sdělením. Jako příklad poslouží včera zveřejněná zpráva o počtu obyvatel ČR k roku 2004. Poslední kategorií jsou aktuální informace s aktuálním sdělením. Tou může být před hodinou zveřejněná předpověď počasí na příští den či výsledek právě skončeného fotbalového utkání.

S jakou kategorií tedy souvisí problém aktuálnosti vyhledávaných informací v Internetu? S první kategorií zcela jistě ne. Na aktuálnosti údajů o Napoleonově tažení do Ruska se již nikdy nic nezmění, takže je nám rovněž lhostejno, kdy byly zveřejněny na Internetu. U druhé kategorie je situace o něco komplikovanější. Neaktuálnost vložení informací nám může být vcelku lhostejná, pokud budou aktuální údaje. Jejich platnost by však měla být uvedena. Pokud tomu tak nebude, pak se může stát, že firmu budeme chtít navštívit, ale na nalezené adrese ji již nezastihneme. Rovněž třetí kategorie není zcela bez problémů. Čerstvé datum uveřejnění je jistě potěšitelné a zvyšuje důvěryhodnost informace. Navíc pokud je u zdroje informace výslovně uvedeno, že se jedná o údaj z roku 2004, pak s tím můžeme počítat a problémy nám to nenadělá. Pokud však údaj o časové platnosti sdělení uveden není, pak předpokládáme, že je informace platná v době, kdy ji čteme. To by v tomto případě však neplatilo. Naši povinností by tedy mělo být ověřit si, zdali tomu tak ještě je. U poslední kategorie je situace jednoduchá. Čerstvá informace s čerstvým obsahem je aktuální, a tedy s největší pravděpodobností i důvěryhodná. Problém aktuálnosti souvisí tedy se druhou a třetí kategorií – tj. s neaktuálními informacemi (ve smyslu data jejich uložení) a neaktuálními údaji. V dalším textu již budu uvažovat jen tyto 2 kategorie.

3.3.3 Příčiny neaktuálních informací

Aktuálnost vyhledávaných informací v Internetu závisí především na dvou faktorech. Prvním jsou tvůrci stránek. Ne každý je dosti pilný, neúnavný a pracovitý, aby aktualizoval své stránky pravidelně. K tomu ho ostatně nemůže nikdo donutit. Druhý problém souvisí s vyhledávacími stroji. Roli „sběrače“ informací u nich hraje robot. Ten shromažďuje WWW dokumenty, ze kterých je pak utvářen index (tj. organizace údajů) [39]. Hlavním cílem robota je, aby shromažďované informace byly co možná nejaktuálnější a nejúplnější. Robot tedy vyhledá dokument a zaindexuje ho. Takto zaindexovaný dokument již může uživatel najít pomocí daného vyhledávače. Jenže Internet je otevřený, a tak v něm přibývají dokumenty stále nové a nové a některé zastaralé zase mizí. Tak se stane, že zaindexované umístění již ve skutečnosti nemusí existovat, tj. index se stává neaktuálním. Jak si s tím vyhledávací systém poradí? Buď se nějakým způsobem indexu ohlásí, že byl dokument změněn či odstraněn a ten ho zaindexuje znovu, nebo je prováděna pravidelná reindexace. Ať již je příčina jakákoli, může nastat situace, kdy je nějaké umístění neaktuální. Tak vznikají tzv. mrtvé odkazy. Většina uživatelů se již jistě setkala s nepopulárním chybou číslo 404 „soubor nenalezen“. Fakt, že vyhledávač najde nefunkční odkaz, nemusí nutně znamenat nezáměr o

aktualizaci dané stránky. Dotyčný se naopak mohl rozhodnout, že zrenovuje své stránky. Za tím účelem je mohl přemístit a vyhledávací stroj zkrátka ještě nezaindexoval nové umístění.

Frekvence indexace určité stránky je závislá na tom, jak často je stránka aktualizována. Svým způsobem je to logické. Vyhledávací stroj nebude reindexovat stránky, na kterých se nic nemění. Při svém vzniku má každá stránka pevně stanovený interval indexace. S častými změnami stránky se interval zkracuje a při neměnném obsahu se naopak prodlužuje.

Čerstvost odkazů závisí také na periodě indexace vyhledávacího stroje. Například vyhledávač Google indexuje nepřetržitě. Reindexace celého Internetu mu trvá cca jeden měsíc. Pokud tedy přemístíte již zaindexovanou stránku, pak hit na ni bude ve výsledcích vyhledávání Googlu aktualizován nejdéle za měsíc. Někomu se to může zdát jako dlouhá doba, ale běžné vyhledávače plně reindexují za 3–4 měsíce. Jedním z nejlepších je v tomto ohledu vyhledávač AlltheWeb, který svůj index plně obnoví během 9–12 dnů [40]. Na druhou stranu pokrytí Internetu AlltheWebem rozhodně nedosahuje rozměrů Google. Menší rozsah se zkrátka reindexuje rychleji.

3.3.4 Jak ovlivnit aktuálnost?

Stejně jako u důvěryhodnosti musíme rozlišit případy z pohledu uživatele a vyhledávacího stroje. Pokud uživatel požaduje co nejaktuálnější dokumenty, pak bohužel nemá moc nástrojů, jak je najít. Větší část odpovědnosti za nalezení aktuálních dokumentů nese vyhledávací stroj.

Jediným přímým nástrojem v rukou uživatelů je filtr data. Většinou se nachází v pokročilých možnostech vyhledávání, které se nám zobrazí po kliknutí na odkazy typu „Pokročilé možnosti vyhledávání“ či „Advanced Search“. Rozsah konkrétních možností se však různí. Například vyhledávač Google umožňuje vyhledat stránky aktualizované za uplynulý rok, čtvrtrok, půlrok či kdykoli. Jiné služby umožňují vyhledávat stránky změněné po nebo před určitým datem (např. Exalead). Filtr zde může být součástí dotazu pomocí příkazů „after“ a „before“. Dotaz ve formátu *after:24/12/2005 fotbal* vyhledá všechny dokumenty, modifikované po loňském Štědrém dnu a obsahujícím slovo fotbal. Obdobně se používá filtr „before“. U některých vyhledávačů lze požadavek aktuálnosti určit až zpětně (např. u Live Search). Většinou to souvisí s možnostmi seřazení nalezených výsledků podle data poslední aktualizace. Mnoho vyhledávačů zobrazuje u popisu nalezených výsledků rovněž poslední datum, kdy stránku prolezli roboti. Na samotné stránce je datum poslední aktualizace obvykle zobrazeno na úvodní stránce webu, většinou v jednom z rohů.

3.3.5 Aktuálnost a vyhledávací stroje

Čerstvost (tj. stáří hitů) vyhledávacích strojů je různá. Ze studie z roku 2003 [41] plyne, že většina hitů je stará cca 1 měsíc. V ní bylo provedeno 6 vyhledávání u 8 vyhledávacích služeb. Zkoumalo se pouze stáří relevantních nálezů, u kterých bylo určeno datum poslední aktualizace. Výsledky jsou uspořádány v tabulce 4.

Vyhledávač	Nejnovější hit	Hrubý průměr	Nejstarší hit
MSN	1 den	4 týdny	51 dnů
HotBot	1 den	4 týdny	51 dnů
Google	2 dny	1 měsíc	165 dnů
AlltheWeb	1 den	1 měsíc	599 dnů
AltaVista	0 dnů	3 měsíce	108 dnů
Gigablast	45 dnů	7 měsíců	381 dnů
Teoma	41 dnů	2,5 měsíce	81 dnů
WiseNut	133 dnů	6 měsíců	183 dnů

tab. 4 - Aktuálnost vyhledávacích strojů

Zajímavé je rovněž srovnání procent mrtvých odkazů u jednotlivých vyhledávačů v roce 2000 [42]. Výsledky jsou již poměrně staré, a tak jen uvedu, že zcela nejhůře dopadl vyhledávač AltaVista s 13,7 % mrtvých hitů. Následovaly služby Excite (8,7 %) a Northern Light (5,7 %). Google měl v roce 2000 4,3 % „slepých ramen“. Navíc srovnání této a několika předešlých studií přineslo zajímavý poznatek. Počet mrtvých odkazů se stále zvětšuje.

4 Vyhledávací stroje a jejich srovnání

Role vyhledávacích strojů je pro vyhledávání informací v Internetu naprosto nezbytná. Není v našich silách sledovat všechny nové informace, které se na Internetu každým dnem objevují. Pokud tedy chce uživatel zjistit o co přišel, může užít vyhledávacích služeb. Nemá-li o nějakém tématu vůbec žádné potuchy, je užití vyhledávače taktéž na místě. Jestliže si chce zkrátka jen obohatit znalosti, vyhledávací služby mu rády vyhoví. Rozsah možností využití vyhledávačů je skutečně široký. Nutno upozornit, že užití vyhledávačů není vždy přínosné. Pokud již známe svá umístění se zaručenými informacemi, pak jistě nemá cenu vyhledávat je pomocí vyhledávačů. Známe-li například adresu nějakého finančního serveru, pak by bylo naprosto zbytečné hledat aktuální kurz koruny k euru pomocí vyhledávače. Vždy než přistoupíme k využití vyhledávače, zamysleme se tedy nad tím, zdali je jeho použití na místě.

Pro popis a srovnání vyhledávačů jsem zvolil 3 zahraniční vyhledávače – Exalead, Google, Windows Live Search – a 2 ryze české vyhledávače – Jyxo a Morfeo. Nejvyužívanější internetové vyhledávače ve světě obsahuje tabulka 5 [1]. Stav je aktuální k červenci 2006. Tabulka nerozlišuje, zda se jedná o vyhledávací stroj či katalog. Sečteme-li procenta podílů na trhu, dostaneme se na číslo 96,6 %. Pro zbylé vyhledávače tedy zbývá pouhopouhých 3,4 %. V nich je obsažen také podíl vyhledávače Exalead. Windows Live Search spustil plný provoz až v září 2006, ale jeho předchůdce MSN je na 3. místě.

Vyhledávač	Podíl na hledání (v %)	Počet hledání/1 návštěvu
Google	43,7	4,4
Yahoo	28,8	4,4
MSN	12,8	3,3
AOL/Time Warner	5,9	2,6
Ask	5,4	3,3

tab. 5 - Podíl vyhledávačů na trhu

Podíl českých vyhledávačů na trhu ukazuje tabulka 6 [2]. Stav je zhruba rok starý (konec listopadu 2005). Pokud vyhledávač slouží zároveň jako katalog i vyhledávací stroj, jsou uvedena procenta součtem za obě využití. Zmíněná studie zjistila, že na českém trhu dominují zaběhnuté portály. Přetrvávající trendem je snižování náskoku Seznamu ze strany Googlu. Podíl vyhledávače Jyxo sice činí necelé 1 %, avšak jeho vyhledávací technologii využívají též portály Atlas a Quick. Celkový podíl Jyxa by tedy měl být spíše 7,6 %. Technologii vyhledávače Morfeo zase využívá portál Centrum, takže by Morfeo měl mít správně 8 %.

Vyhledávač	Podíl na trhu (v %)
Seznam.cz	47,7
Google	32,9
Atlas	6,9
Centrum	5,2
MSN	1,5
Jyxo.cz	0,7
Navrcholu.cz	0,7
Zoohoo.cz	0,7

tab. 6 - Podíl vyhledávačů na českém trhu

U následujícího srovnání vyhledávacích služeb nebudu provádět multikriteriální výběr. Všechny probírané vyhledávače se liší zejména ve specializaci na vyhledávané oblasti, relevanci výsledků, podporovaných operátorech a také vzhledu rozhraní. Hlavně relevance výsledků a vzhled rozhraní jsou záležitosti čistě subjektivní. Jakékoli kvantifikované srovnávání by tedy porovnávalo neporovnatelné. Pro celkový přehled vyhledávačů a jejich funkcí doporučuji stránku <http://www.internettutorials.net/choose.html>, která také poradí, jaký vyhledávač zvolit v závislosti na povaze hledaných informací.

4.1 Exalead

4.1.1 Obecné údaje

Exalead je francouzský vyhledávač nové generace, který spustil činnost v říjnu roku 2004. Je součástí ambiciózního projektu Quaero, jehož se účastní i giganti z jiných oborů jako například Thomson, Siemens AG, France Telecom a další [3]. Cílem tohoto projektu je rozvoj nových technik v oblasti vyhledávání, kterými se Exalead odlišuje od ostatních vyhledávacích strojů. Již samotný název Exalead naznačuje leccos z jeho hlavních cílů. Předpona exa má zdůraznit jeho obrovský rozsah pokrytí webu (exabyte je 10^{18} bytů) a sloveso lead (vést) značí pomoc uživatelům při vyhledávání, kdy je tazatel doslova veden vyhledávačem k účinnému nalezení informací. Velikost indexu Exaleadu činí v současné době něco málo přes 8 miliard stránek. Exalead je špičkou ve vyhledávání multimediálních informací jako jsou videa a zvukové záznamy. Poskytuje však rovněž vyhledávání obyčejných textově orientovaných dokumentů. Exalead je především vyhledávací stroj, ale svou databázi též synchronizuje s hierarchií adresářů Open Directory. Jedná se o projekt dobrovolníků, spravujících jakýsi celosvětový katalog. Projekt je podporován firmou Netscape. Adresáři Exaleadu však nelze nijak procházet (alespoň jsem na to nepřišel), jak je tomu např. u Googlu. Vyhledávač je k dispozici na adrese www.exalead.com.

4.1.2 Konkrétní vlastnosti

Jako vůbec první nás na úvodní stránce Exaleadu upoutá jeho uživatelské rozhraní (viz obrázek 10 na následující straně). Exalead má pro mne opravdu nejpříjemnější design, s jakým jsem se kdy u vyhledávačů setkal. Není tam nic navíc a nic tam také nechybí. V pravém horním rohu se nacházejí odkazy „Sign in“ a „Preferences“. Tlačítko „Sign in“ slouží k přihlášení ke svému účtu. Účet poskytuje emailovou schránku a možnost ukládat své vlastní prostředí a vzhled obrazovky, které jsou nastavitelné v předvolbách „Preferences“. Jejich možnosti proberu později. Jen předešlu důležitou volbu nastavení jazyka rozhraní. K dispozici jsou angličtina, francouzština, němčina, italština a španělština. Čeština tedy zatím chybí, což by mohlo u mnohých českých uživatelů hrát velkou roli. Uprostřed úvodní stránky nad polem pro zadávání dotazů je ikona Exalead, sloužící k navrácení na úvodní stránku vyhledávače. Ta je viditelná v každé fázi vyhledávání. Pod ní se nachází karty „Web“ a „Images“. Karta „Web“ slouží ke klasickému vyhledávání, karta „Images“ k vyhledávání obrázků. Jako tlačítko spouštějící vyhledávání slouží „Web Search“, popř. „Image Search“. Vedle něho se nachází odkaz „Advanced search“, které uživatele posune do pokročilých nastavení vyhledávání. O nich ale až později. Pod polem pro dotaz se nachází odkaz „create an account“, sloužící k založení účtu. Pod ním se nachází prostor s ikonami, které může uživatel využít jako rychlý přechod na přednastavené oblíbené stránky. Považuji je za velice chytře využitě místo. Můžete si zde nastavit své oblíbené stránky jako záložky. Navíc jsou zobrazeny i s miniaturami jejich úvodních stran. Umístění záložek se ukládá k vašemu účtu nebo do vašeho počítače. Pokud tedy spustíte Exalead z jiného počítače, pak záložky nebudou bez přihlášení k účtu k dispozici. Aby byl výčet položek na úvodní straně kompletní, zmíním ještě odkazy „Add more shortcuts“ pro přidání dalších záložek a k vyhledávání nepotřebné odkazy s nabídkami prací, informacemi o Exaleadu a možností odeslat připomínky ohledně vyhledávače. Tolik k úvodní straně Exaleadu.

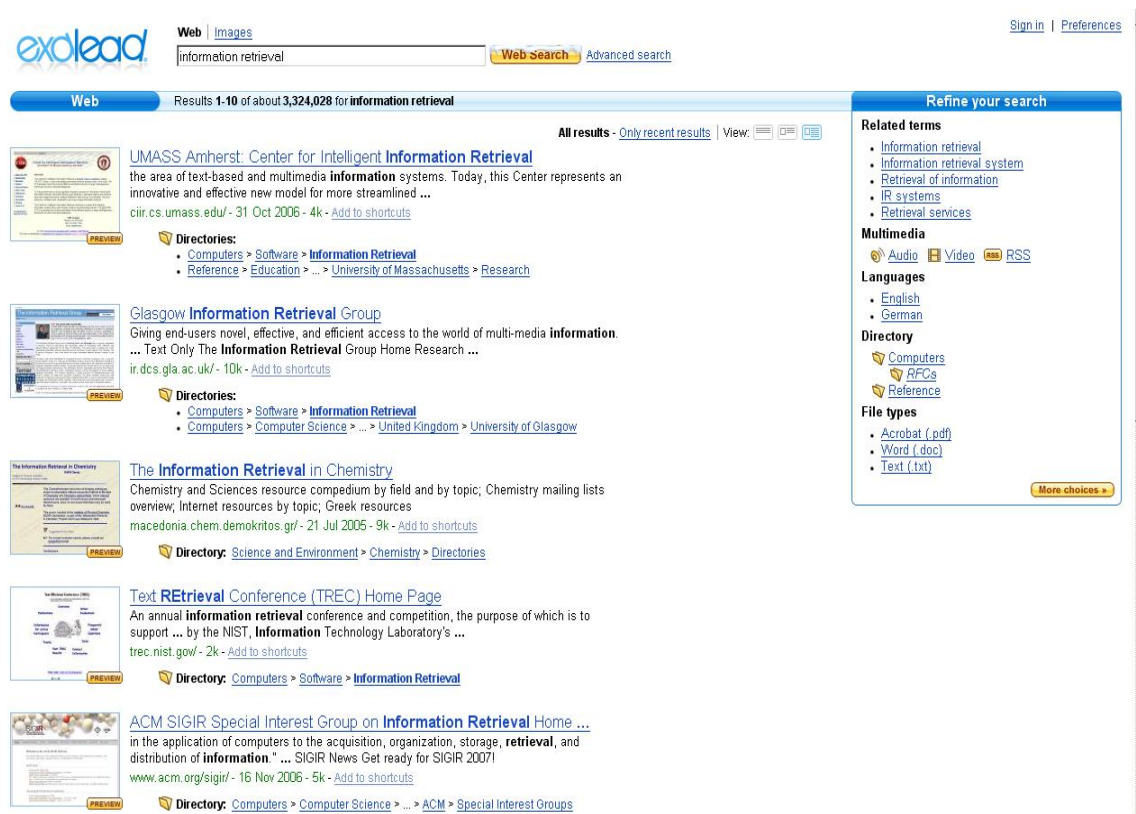


obr. 10 - Úvodní strana vyhledávače Exalead

Prostředky Exaleadu a syntaxe zadávání dotazu jsou o něco bohatší než u jiných vyhledávačů. Podporovány jsou booleovské operátory AND, OR, NOT, symbol „–“, užívání závorek a nejtypičtější filtry. Navíc je zde operátor OPT, vyjadřující nepovinnost výskytu výrazu ve výsledcích. Jako u jednoho z mála vyhledávačů je podporován zástupný znak „*“, sloužící ke zkracování slov. Pro dvě a více slov používá Exalead stejně jako většina vyhledávačů zkrácený zápis slov automaticky (tzv. automatic stemming). Standardním operátorem je AND. Z proximitních operátorů lze využívat operátoru NEAR, NEXT a frázi. U operátoru NEAR lze navíc zvolit vzdálenost hledaných termínů (ve formátu NEAR/#). Operátor NEXT vyjadřuje požadavek, aby se ve výsledku vyskytovala klíčová slova vedle sebe, přičemž na jejich pořadí nezáleží. Všechny operátory musí být psány velkými písmeny. Právě u nich jako jediných záleží na velikosti písmen. Stopslova jsou běžná jako u všech vyhledávačů. K vynucení jejich výskytu slouží symbol „+“.

Nyní se dostávám k samotným funkcím vyhledávače. V dalším textu budu předpokládat prvotní dotaz *information retrieval*. Exalead obohacuje výsledky vyhledávání o několik velmi užitečných informací a statistik (viz obrázek 11 na další straně). Jedinečnými doplňky jsou miniatury úvodních stránek, které byly nalezeny, geografické informace a informace o kategoriích výsledků. Hlavně miniatury (tzv. thumbnails) jsou velice užitečné. Ihned podle nich poznáte, jaký má nalezená stránka design. A to je přece nejdůležitější faktor důvěryhodnosti stránky. Někomu se sice může zdát, že jsou náhledy příliš malé, ale podle mne stačí. Pokud se nezamlouvají, je možné přepnout do klasické struktury výsledků bez miniatur. Kliknete-li na miniaturu stránky, zobrazí se její podoba při poslední indexaci. Stejnou možnost poskytuje i Google. Počet nalezených výsledků, umístěný pod polem pro dotaz, je samozřejmou součástí každé obrazovky s výsledky. To samé lze říci o titulku nalezené stránky (název hitu), URL adrese, několika prvních slovech stránky, její velikosti a datu poslední indexace. Neobvyklé je však umístění nalezené stránky v hierarchii

adresářů Exaleadu, pokud je tam však stránka zařazena. Pro můj dotaz byla například nalezena stránka, jejíž místo v hierarchii bylo: Computers -> Software -> Information Retrieval.



obr. 11 - Stránka s výsledky Exaleadu

Neobvyklý je také rámeček v pravé části obrazovky. Považuji ho za perfektní nápad. Obsahuje totiž nástroje ke zjemnění vyhledávání, přičemž však nemusíte opouštět stránku s výsledky. Využit lze hledání příbuzných termínů a multimediálních souborů k původnímu dotazu, jiného jazyku vyhledávání a jiných adresářů v katalogu a vyhledávání podle typu souborů a geografického umístění. Ale popořadě. Příbuzné termíny v mém případě zahrnovaly pojmy „Information retrieval system“, „Retrieval of information“ aj. Nástroj je tedy výhodný tehdy, pokud hledáme něco konkrétního a nejsme si jisti svým původním dotazem. Multimediální typy souborů zahrnují audio a video soubory a také RSS zdroj (tzv. RSS feed). RSS je typ XML souboru, který je určen ke čtení novinek na webových stránkách. RSS formát poskytuje obsah celého článku, popř. jeho část [4]. Možnost vyhledávat stránky v určitém jazyce je dnes dostupná u většiny vyhledávačů. Exalead však podporuje vůbec nejvíce jazyků ze všech vyhledávačů – 54 (stav k říjnu 2006). Čeština mezi nimi samozřejmě figuruje. Pokud pouze zpřesňujeme dotaz, jsou v rámečku k dispozici jen ty jazyky, ve kterých vyhledávač našel nějaké výsledky při prvním vyhledání. Je dokonce uveden jejich procentní výskyt. Na můj anglický dotaz *information retrieval* nabízí Exalead zúžení dotazu do 6 jazyků. První je samozřejmě angličtina s 97 %. Vyhledávání v jiných adresářích katalogu umožňuje specifikovat oblast, ve které se má původní dotaz vyskytovat. K dispozici jsou v tomto případě adresáře „Reference“, „Society and Culture“, „Computers -> RFCs“ a další. Možnost vyhledat jen určité typy souborů je taktéž velmi bohatá. Kromě obvyklých přípon .pdf (Acrobat), .doc (Word) a .txt (Text) jsou na výběr přípony .ppt (PowerPoint), .rtf (Rich Text Format), .xls (Excel), .swf (Flash) a .wpd (Word Pad). Vyhledávání podle geografické lokace dokumentu je podobně jako u jazyka závislé na prvním vyhledání, tj. nabídky závisí na umístění dokumentů, nalezených původním dotazem. Celkem 71 % všech dokumentů z prvního hledání se v mém případě vyskytovalo v Severní Americe, kde lze

specifikovat další 3 země. Poslední možností rámečku je hledání uvnitř výsledků, což je ale dostupné u každého lepšího vyhledávače. U vyhledávání obrázků umožňuje rámeček specifikovat jeho velikost (malý, střední, velký), rozlišení, barvu (barevný, černobílý), rozvržení (krajina, portrét) a typ souboru (JPEG, GIF, PNG).

Již jsem zmínil odkaz předvoleb „Preferences“. Kromě volby jazyka rozhraní lze napevno nastavit jazyk vyhledávání, tedy jeden z oněch 54. Užitečnou možností hlavně pro rodiče je aktivace a deaktivace filtru pro děti. Můžete ho vypnout, zapnout či nastavit filtr mírný (tj. filtr zapnutý jen při vyhledávání obrázků). Další možnosti jsou již zanedbatelné a na nalezené výsledky nemají vliv. Souvisí především se způsobem jejich zobrazení (kolik výsledků na stránce u běžného vyhledávání, kolik u obrázků apod.).

Možnosti rozšířeného vyhledávání obsahují oproti zúžovacímu rámečku jen pár nástrojů, které lze použít již před zadáním prvního dotazu. Souvisí především s booleovskými operátory, zkracováním slov pomocí zástupného znaku „*“ a užitím nejtypičtějších filtrů. Dva z filtrů jsou však neobvyklé. Jedná se o filtry „soundslike“ a „spelllike“. Filtr „soundslike“ slouží pro případ hledání pojmu, u kterého známe pouze jeho fonetickou výslovnost. Problémem je, že v každém jazyce slyší lidé trochu jinak. Pokud tedy zadáte dotaz *soundslike:informeješn ritrívl*, nečekejte, že budou nalezeny dokumenty o pojmu „information retrieval“. Osobně jsem s tímto filtrem neměl dobré zkušenosti ani u anglických výrazů. Výsledky při použití filtru byly víceméně totožné s výsledky bez filtru. Filtr „spelllike“ je podobný. Jen s tím rozdílem, že má pomáhat při pátrání po slovech, u kterých neznáme přesný pravopis. Ani tento filtr mne však nepřesvědčil, ačkoli byl užitečnější než filtr „soundslike“.

Velmi důležitým pomocníkem jsou regulární výrazy. Regulární výraz je šablona, která popisuje sadu řetězců [5]. Obvykle se používá ke stručnému popisu této sady, aniž bychom museli vypisovat všechny její prvky. Regulární výraz je v Exaleadu ohraničen symboly „/“ („lomítka“). Uvnitř něj reprezentuje symbol „.“ (tečka) jeden libovolný znak, symbol „*“ (hvězdička) opakovací znak, symbol „|“ (svislice) má funkci OR a v obvyklém významu se užívají závorky. Symbol „?“ (otazník) slouží k vyjádření nepovinnosti znaku. Například dotaz */ex(p|t|c)..s*..n?/* najde pojmy „experts“, „extreme“, „extension“, „expansion“, „excursion“ apod.

Za výtečnou považuji rovněž možnost určit si pořadí výsledků vyhledávání. Standardně jsou výsledky řazeny dle relevance. K řazení dle stáří slouží operátor „sort“. K dispozici jsou 2 možnosti hodnot – „old“ a „new“. „Old“ řadí výsledky vzestupně od nejstarších, „new“ naopak. Na dotaz *sort:old information retrieval* je na 1. místě stránka z roku 1972. K dispozici jsou též filtry „before“ a „after“, o jejichž použití již byla řeč v kapitole věnované aktuálnosti.

4.1.3 Shrnutí

Jako nedostatky jsou Exaleadu vytýkány nižší relevance výsledků ve srovnání s top vyhledávači, menší databáze stránek, relativní neznámost v podvědomí veřejnosti a delší prodleva při reindexaci. Přidal bych ještě pomalost vyhledávání, hlavně co se týče obrázků. Někomu by též mohla vadit absence češtiny jako možného jazyka rozhraní. Mezi výhody naopak patří široká podpora operátorů (asi vůbec nejširší), miniatury nalezených odkazů a zúžovací rámeček na pravé straně obrazovky. Pro mě je Exalead také jedničkou ve vzhledu uživatelského rozhraní. Možnosti rozšířeného vyhledávání považuji za velmi slušné a jeho náповědu za vůbec nejnázornější. Málokterý vyhledávač umožňuje řadit výsledky podle stáří, Exalead však ano.

4.2 Google

4.2.1 Obecné údaje

Google není jen vyhledávačem, je rovněž obrovskou korporací, jednou z nejbohatších na světě, která po celé planetě zaměstnává tisíce lidí. Slovo Google je dnes synonymem volného přístupu k informacím a také velmi cennou obchodní značkou, stejně jako například Mc'Donalds či

Coca-Cola. Vždyť i samotný proces vyhledávání se někdy, nutno říci mylně, označuje pojmem „googlování“. Pokud by se tedy kvalita vyhledávače měřila na světový věhlas, byl by Google jasně nejlepším. Jenže právě věhlas Googlu je tím, co ho činí nepoužívanější vyhledávací službou na světě. Jsem přesvědčen, že mnoho lidí užívá Googlu právě proto, že je nejpoužívanější, aniž by tušili, v čem tkví jeho skutečné přednosti. Přesto si neumím představit, že by ho v popularitě mohl v příštích letech nějaký vyhledávač předstihnout.

Vyhledávač Google spustil činnost v září 1999 a v současné době (resp. k červenci 2006) je primárním vyhledávačem pro 43,7 % uživatelů Internetu. Každodenně obslouží přes 200 milionů dotazů [6]. Velikost indexu Googlu je odhadována na závratných 25 miliard stránek a 1,3 miliard obrázků. Navíc Google prochází i stránky, které dosud nemá zaindexovány. Také Google je součástí zmíněného Open Directory Projectu, což z něj dělá částečně i katalog. Rozsah oblastí, ve kterých Google dokáže vyhledávat, je nejširší ze všech vyhledávačů. Kromě základních oblastí jako jsou webové stránky a obrázky umožňuje prohledávat také databázi video souborů, zpravodajských novinek, diskusních fór, knih, map, internetových obchodů, webových logů a nejnověji také patentů. Vyhledávač je dostupný na adrese www.google.com, jeho česká verze na www.google.cz.

4.2.2 Konkrétní vlastnosti

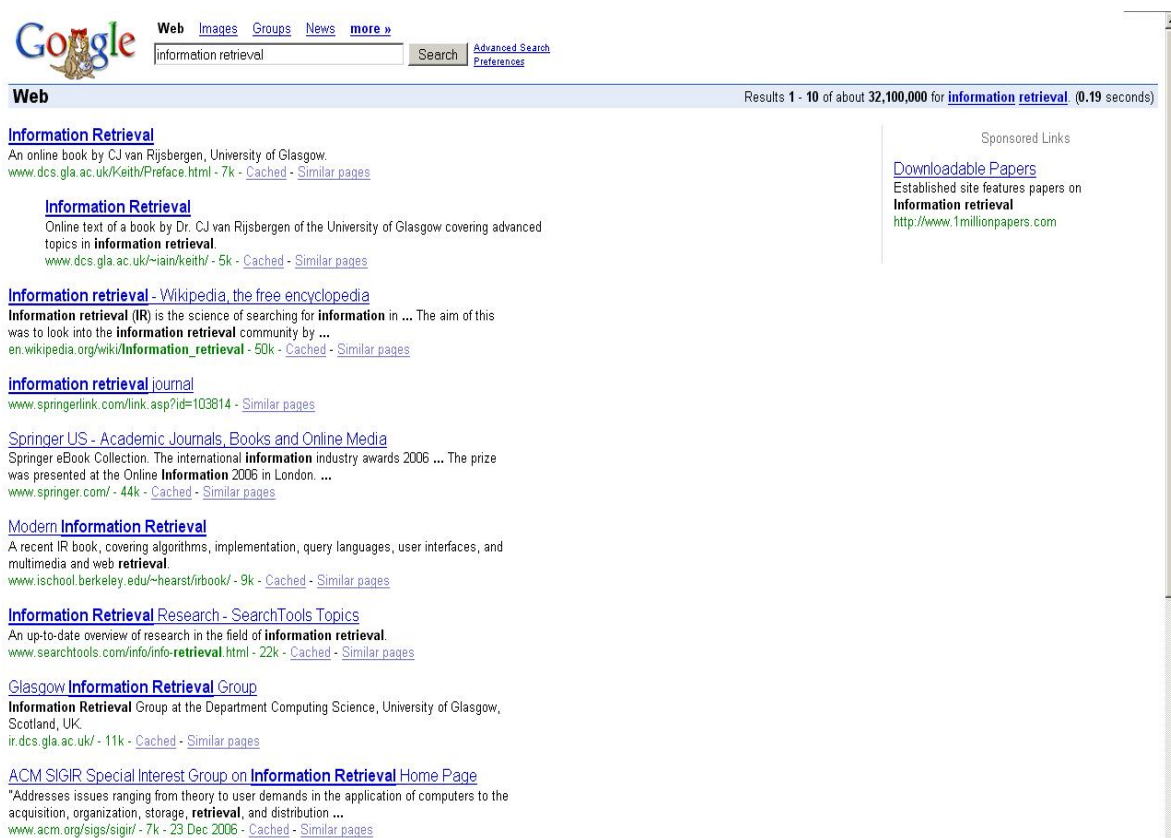


obr. 12 - Úvodní stránka Googlu

Uživatelské rozhraní Googlu (viz obrázek 12) je velice přehledné, podle mne však nikoli pohledné. Záhlaví obrazovky vévodí mocné logo Google, které je vždy připraveno přesunout uživatele na úvodní stranu. V následujícím textu budu k anglickým názvům tlačítek do závorek uvádět jejich ekvivalenty v české verzi Googlu. Na úvodní stránce je jediné vstupní pole pro dotaz, který se odesílá pomocí tlačítka „Google Search“ („Vyhledat Googlem“). Vedle něho je k dispozici tlačítko „I’m Feeling Lucky“ („Zkusím štěstí“). Jeho využití mi nepřijde moc výhodné. Místo klasické stránky s výsledky vás nasměruje rovnou na stránku, která by ve výsledcích odpovídala nejrelevantnějšímu nálezu. Málokdy se však stane, že by tou nejuspokojivější stránkou byla právě ta první. Navíc se při vyhledávání většinou nespokojím jen s jedním hitem. Nad vstupním polem se nachází odkazy „Web“ („Web“), „Images“ („Obrázky“), „Groups“ („Skupiny“) a „News“ („Novinky“). První dva mají stejné funkce jako u Exaleadu. „Groups“ slouží k vyhledávání v diskusních skupinách a „News“ prochází seznam zpráv, které nejsou starší než 1 měsíc. Pod odkazem „more“ se nachází zbývající oblasti. Napravo od dotazového pole se nachází 3 důležité

odkazy. „Advanced Search“ („Pokročilé vyhledávání“) nás nasměruje do pokročilých funkcí vyhledávání, „Preferences“ („Nastavení“) umožňuje nastavit několik voleb včetně jazyka a „Language Tools“ („Výběr jazyka“) povoluje nastavit některý z mnoha jazyků či překlad textu z některých jazyků do jiných. Vše vysvětlím v dalších odstavcích. Výčet prvků úvodní strany zkompletuji nedůležitými odkazy na reklamní programy a informace o Googlu.

Pravidla pro zadávání dotazů jsou stejná jako u většiny vyhledávačů. V dotazech se nerozlišuje velikost písmen. Jedinou výjimkou je slovo „or“. Pokud ho chcete použít jako booleovský operátor, pak ho musíte napsat velkými písmeny. OR je také spolu se symbolem „-“ jediným podporovaným booleovským operátorem. Operátor AND doplňuje Google automaticky. Z proximitních operátorů jsou povoleny jen fráze, které Google rozpoznává i bez použití uvozovek. Jediným podporovaným zástupným znakem je „*“, ale nelze ho užít ke zkracování slov. V Googlu je určen k zastoupení celého slova (někdy se stává, že jím Google zastoupí i více slov). Symbolu lze užít i uvnitř fráze. Google stejně jako většina vyhledávačů používá automaticky zkrácený zápis slov (tzv. automatic stemming), který je v podstatě náhražkou zástupných symbolů. Někdy však místo pomoci naopak přitíží. Google totiž uživateli nesdělí, u kterých slov stemming provedl a u kterých ne. Škála stopslov je obvyklá. Stopslova budou brána v potaz, pokud se vyskytnou v uvozovkách či jako součást fráze, nebo pokud před ně umístíme symbol „+“ (plus). Mezi symbolem a slovem však nesmí být mezera. Limit pro dotaz je 10 slov. Pokud zadáme dotaz delší, Google nám ho automaticky ořízne o přečnívající slova tak, aby měl dotaz 10 slov. Zmíním také symbol „~“ (tilda), který se užívá k nalezení synonym. V současné době však funguje pouze u anglických slov.

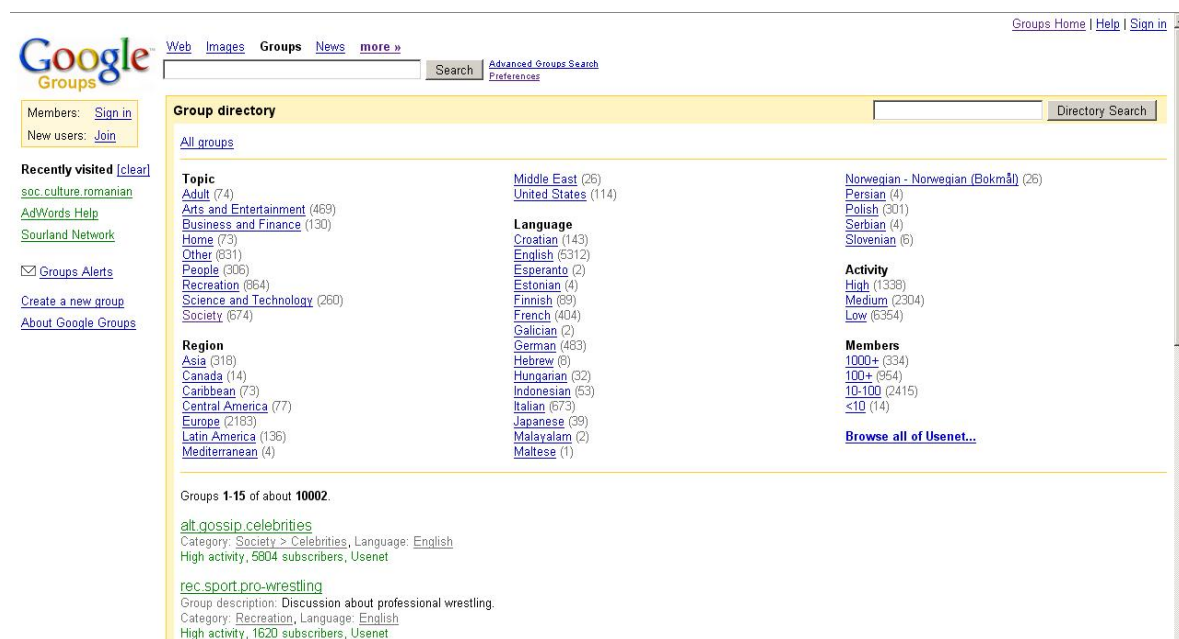


obr. 13 - Stránka s výsledky Googlu

Stránka s výsledky (viz obrázek 13) není odlišná od jiných vyhledávačů. Nad textovým polem pro dotaz se nachází stejné odkazy jako na úvodní stránce. Kliknete-li na některý z nich, dotaz se odešle jako hledání obrázku (resp. ve skupinách, resp. ve zprávách), aniž by jste ho museli psát znovu. Pod polem pro dotaz se nachází řádek s údaji o počtu nalezených výsledků, zněním

dotazu a rychlostí vyhledání. Právě ta je u Googlu velmi působivá, a to i v dotazech vedoucích na miliony hitů. U každého hitu se vypisují obligátní položky – název, několik prvních řádků obsahu, URL adresa a velikost stránky. Datum poslední indexace je uvedeno pouze v případě, pokud byla stránka indexována nedávno. Neobvyklé jsou však dva odkazy. „Cached“ („Archiv“) zobrazí stránku tak, jak vypadala, když ji Google prošel naposledy a „Similar Pages“ („Podobné stránky“) odkazuje na stránky s podobným obsahem. Navíc pokud je stránka v jiné řeči, než je náš jazyk nastavený v „Preferences“ a Google podporuje překlad do tohoto jazyka, pak se vedle názvu hitu objeví odkaz „Translate this page“ („Přeložit tuto stránku“). Po využití této možnosti si můžeme přečíst obsah stránky v přednastaveném jazyce. Nutno podotknout, že překlad do češtiny není podporován. Při nalezení jiných typů souborů (např. .pdf) se objeví odkaz „View as HTML“ („Zobrazit jako HTML“), který stránku převede do podoby HTML.

Již jsem zmínil možnost vyhledávání obrázků, ke kterému se dostaneme kliknutím na odkaz „Images“ („Obrázky“). Odkaz umožňuje vyhledávat ve více než 1,3 miliardách grafických souborů, v čemž je Google bezesporu nejlepší. Google po zadání dotazu hledá nejen v názvu souboru, ale i v titulku, textu obklopujícím obrázek a pokouší se (často neúspěšně) vrátit seznam obrázků bez duplicit. Vyhledávací syntaxe i záhlaví stránky s výsledky jsou v podstatě stejné jako v sekci Web. Výhodný je řádek „Show“ („Ukaž“), který umožňuje specifikovat velikost obrázků. Výchozí volba je „All sizes“ („Libovolně velké“). Dalšími možnostmi jsou velké, střední a malé obrázky. Každý obrázek je zobrazován jako miniatura s údaji o názvu souboru, původním rozlišení, velikosti a URL adrese.



obr. 14 - Adresář skupin Googlu

„Groups“ („Skupiny“) považují za velkou přednost Googlu oproti svým konkurentům. Vyhledávání ve skupinách pracuje s databází nejstaršího z internetových diskusních fór, USENETu [7]. Tisíce uživatelů zde dennodenně publikují své příspěvky. Uživatelé Googlu se tak naskytuje jedinečná možnost velmi jednoduchého prohledávání celého archivu zpráv USENETu vystavených od roku 1995. V současné době jich je okolo 1 miliardy. Pokud tedy hledáte řešení nějakého problému, můžete se se svým dotazem obrátit na skupiny Googlu. Obrací se sem i profesionálové ze všech možných oborů. Hlavním rozdílem od vyhledávání v sekci Web jsou tématické skupinové odkazy (viz obrázek 14). K dispozici jsou např. skupiny „Arts and Entertainment“ („Umění a zábava“), „People“ („Lidé“), „Business and Finance“ („Obchod a finance“) aj. Vyhledávat lze též podle regionu a jazyka zpráv. Jednotlivé zprávy ve skupinách jsou uspořádány podle toku

konverzace. Dvěma nejdůležitějšími operátory pro vyhledávání ve skupinách jsou „author“ a „group“. „Author“ vyhledává zprávy vystavené na fóru. Parametrem může být buď jméno autora, nebo jeho emailová adresa. „Group“ vyhledává v titulcích vystavených zpráv nebo v klíčových slovech popisujících skupinu.

Nastavit Google dle svých možností lze přes odkaz „Preferences“. Jako u Exaleadu lze volit jazyk rozhraní a jazyk hledání. U obou nechybí čeština. Hledat lze v 36 jazycích a je možné zaškrtnout více jazyků najednou. Standardně hledá Google ve všech jazycích. Další shodu s Exaleadem tvoří možnost zapnutí filtru pro blokování sexuálního obsahu nevhodného pro děti a nastavení počtu výsledků zobrazených na 1 stránce s výsledky. Maximem je 100 hitů na stránce. Na stránce jazykových nástrojů „Language Tools“ lze kromě nastavení jazyka a geografického umístění užít také nástrojů pro překlad. K dispozici je formulář pro vložení textu nebo pole pro URL adresu stránky, která se má přeložit. Rozsah jazyků je následující – z angličtiny do 8 jazyků, do angličtiny ze 7 jazyků, z francouzštiny do němčiny a naopak.

Většinu z možností v pokročilém vyhledávání („Advanced Search“) lze vyjádřit pomocí základních operátorů a filtrů, takže v tomto ohledu nenabízejí téměř nic navíc. Za zmínku stojí možnost vyhledávat termíny kdekoli na stránce, v názvu stránky, v URL adrese, těle stránky či v odkazech na stránku (i když i to lze vyjádřit operátory). Užitečná je též možnost vyhledat stránky aktualizované kdykoli, za poslední rok, půlrok či čtvrtrok. Standardem je prohledávání nebo naopak vyloučení určité domény z vyhledávání.

Nyní se dostávám k pokročilým operátorům Googlu. Všechny typické z konce 3. kapitoly jsou podporovány. Je však třeba dávat pozor na malé odlišnosti v použití v závislosti na tom, kde vyhledáváme. Např. filtr „intitle“ vyhledává v titulku vystavené zprávy, pokud vyhledáváme v sekci skupin. Filtr „filetype“ prohledává nepřeberné množství přípon. Oproti Exaleadu jsou navíc důležité přípony .ps (Adobe PostScript), .wks (Microsoft Works), soubory aplikace Lotus 1-2-3 a mnoho dalších. Zajímavými filtry navíc jsou „inanchor“, „numrange“ a „related“. „Inanchor“ hledá textovou reprezentaci odkazu a nikoli skutečnou URL (jako je tomu u operátoru „link“). Dotaz *inanchor:click* vyhledá tisíce stránek, obsahujících odkazy jako „click here“, „click to start“ apod. Operátor „numrange“ slouží k vyhledávání čísel v rozsahu stanoveném horní a dolní mezí. Dotaz *numrange:1-3* vyhledá čísla 1, 2 a 3 kdekoli na stránce. Výhodou je, že přitom ignoruje okolní symboly jako jsou označení měny a čárky. Použít lze též zápisu *1..3*, bez užití samotného operátoru. Operátor „related“ je určen k vyhledání těch webů, které Google určil jako jemu podobné. Dotaz *related:www.vse.cz* tak najde stránky Univerzity Karlovy a dalších škol. Ekvivalentní je užití odkazu „Similar pages“ ve výsledcích vyhledávání.

Google též podporuje několik dalších operátorů. Velmi zajímavými jsou operátory „info“, „define“ a „phonebook“. Operátor „info“ zobrazuje souhrnné informace o webu a odkazy na jiná hledání, které by se ho mohly týkat (např. záznam o stránce v archivu, stránky nacházející se na dané doméně aj.). Pokud vložíme URL adresu jako dotaz, operátor nemusíme psát. „Define“ vyhledá anglickou definici termínu. Vyhledávat lze též v telefonních seznamech (jen po USA), a to pomocí „phonebook“. Pokud uvedete jako dotaz nějakou adresu (nebo alespoň něco, co se jí podobá), pak vám Google automaticky vyhledá mapu dané lokality.

Výsledky v Googlu jsou standardně řazeny dle relevance (zmiňovaný algoritmus PageRank) a seskupovány podle webových stránek. To znamená, že z žádné stránky nebudou ve výsledcích více než 2 hity.

4.2.3 Shrnutí

Výhody Googlu jsou především velikost databáze stran i obrázků, vysoká relevance výsledků, vedený archiv stránek, fascinující rychlost vyhledávání a značný rozsah dalších operátorů. Množství přípon souborů, které lze vyhledávat, je taktéž pestré. Další výhodou je, že pokud uděláte syntaktickou chybu, Google vám ji oznámí. Slabinami jsou absence zkracování slov, neúplná podpora booleovských operátorů a také nemožnost jejich vkládání do sebe. Dalšími

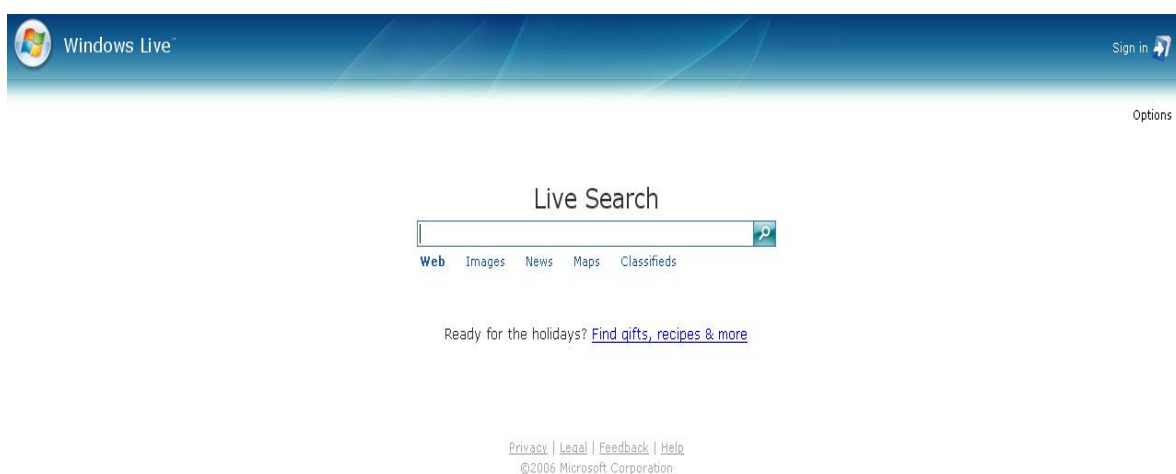
stinnými stránkami jsou ne vždy dokonalé rozpoznání správného typu souboru (to se týká i nejpoužívanějších přípon) a jistá nedbalost při automatickém používání množných čísel a synonym ke klíčovým slovům. Google vám to totiž ani neoznámí.

4.3 Windows Live Search

4.3.1 Obecné údaje

Windows Live Search je stále ještě novým vyhledávačem z dílny Microsoftu. Činnost spustil teprve 11. září 2006 jako následník služby MSN [8]. Transformace proběhla díky snaze Microsoftu zatraktivnit poskytované služby a konkurovat více službám Google a Yahoo!. Live Search používá své vlastní unikátní databáze stránek. Velikost jeho indexu jsem však nikde nenašel. Vyhledávač je dostupný na adrese www.live.com.

4.3.2 Konkrétní vlastnosti

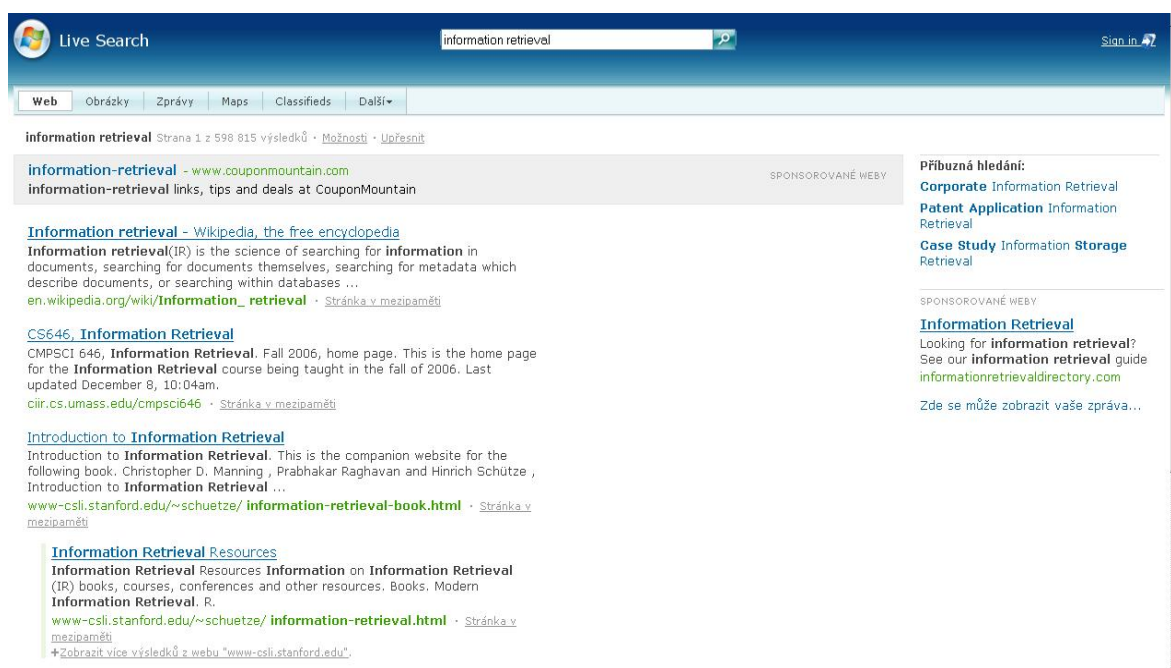


obr. 15 - Úvodní stránka Live Search

Úvodní stránka vyhledávače je velmi jednoduchá (viz obrázek 15). Rovněž Live Search má podle mne příjemnější uživatelské rozhraní než Google. Odkazy na úvodní straně se liší v závislosti na jazyce rozhraní. Pro českého uživatele je jistě příjemné, že vyhledávač podporuje češtinu. Osobně ji však nedoporučuji, poněvadž je české rozhraní chudší než anglické. Uživateli by pak na první pohled nemuselo být zřejmé, kde všude lze vyhledávat. Horním okrajem obrazovky prostupuje modrý pruh, který je spolu s logem také jediným grafickým prvkem celého vyhledávače. Logo se nachází na jeho levém okraji a po jeho přjetí myší se zobrazí odkazy na další sekce a partnerské stránky vyhledávače (MSN, Messenger, Live Mail apod.). V pravém rohu se nachází odkaz „Sign in“ („Přihlásit“), sloužící k přihlášení ke svému účtu. Pod ním se již na bílém pozadí nachází odkaz „Options“ („Možnosti“). Ten slouží k obligátním nastavením jazyka rozhraní a vyhledávání (38 jazyků), rodičovského filtru, obrazovky výsledků (užitečné je nastavit maximální množství hitů na jeden web) a také vaší přesné lokace. S touto možností specifikace jsem se setkal poprvé. Vyhledávač nám dává na vědomí, že mu její určení často dopomůže k vyhledání relevantnějších výsledků. Já mám standardně nastaveno „praha, praha“. Zhruba ve prostředku stránky je textové pole pro dotaz, který se odesílá ikonou lupy, nacházející se napravo od pole. Pod polem jsou odkazy „Web“, „Images“, „News“, „Maps“ a „Classifieds“, které jsou však k dispozici jen v anglické verzi. Česky je najdete až po prvotním vyhledání. Význam prvních čtyř z nich je stejný jako u Googlu a „Classifieds“ slouží k prohledávání adresářů. Informace o ochraně osobních údajů, právní informace, prostor k umístění vlastních názorů a nápověda jsou z hlediska vyhledávání druhotné, avšak lze je využít.

Nyní něco málo k nástrojům vyhledávání. Standardním operátorem je AND. Booleovské operátory jsou podporovány všechny a musí být při svém použití psány velkými písmeny. Seskupování operátorů umožňují standardní závorky. Užití symbolů „-“ a „+“ je obvyklé. Z proximitních operátorů jsou povoleny jen fráze. Zkracování slov či užití zástupných symbolů je nepřipustné. Live Search není stejně jako většina vyhledávačů citlivý na rozlišování velkých a malých písmen. Škála stopslov je stejná jako u jiných vyhledávačů.

Stránka s výsledky obsahuje obvyklé informace (viz obrázek 16) – znění dotazu, další možné oblasti hledání, počet nalezených výsledků, název stránky, několik prvních slov, URL adresu a odkaz na archivovanou podobu stránky. Velikost stránky chybí a datum poslední indexace je vypsáno jen tehdy, pokud byla stránka indexována nedávno. Na pravé straně se vyskytují příbuzná hledání, kterými lze upřesnit původní dotaz.



obr. 16 - Stránka s výsledky Live Search

Pokročilé možnosti vyhledávání jsou k dispozici až po prvotním vyhledání dotazu a zpravidla k němu něco přidávají, tj. slouží jako zužovací rámeček u Exaleadu. Pokročilé možnosti nelze vyvolat ještě před zadáním dotazu, což je poněkud podivné. Objeví se nám formou vysunovacího menu po kliknutí na odkaz „Advanced“ („Upřesnit“) na stránce s výsledky. Možností je celkem 6. K původnímu dotazu můžeme přidat další termíny a určit, aby výsledek obsahoval všechny termíny, libovolné z termínů, přesně přidanou frázi nebo žádný z termínů. Jinými slovy se jedná o rozšíření původního dotazu pomocí booleovských operátorů a frází. Dalšími nástroji pokročilého vyhledávání jsou prohledávání stran v určité doméně, stran s odkazem na určitou URL adresu, stran v určitém geografickém umístění či v určeném jazyce. Jinými slovy se jedná o filtry umístění, jazyka stránky a o operátory „site“ a „link“. Live Search v tomto ohledu nenabízí tedy nic nového.

Velmi užitečná je však možnost „Results ranking“ („Hodnocení výsledků“). Standardně řadí Live Search výsledky dle relevance. Řazení však lze změnit. Jedná se o to, že výsledky určené daným dotazem můžete posunem jezdců seřadit dle aktuálnosti, oblíbenosti a úrovně shody. U aktuálnosti tvoří opačné extrémy stránky nedávno aktualizované a stránky statické. Zde se pracuje s datem jejich vložení do indexu, popř. datem poslední reindexace. U oblíbenosti je možno umístit na čelo výsledků stránky velmi oblíbené či naopak méně oblíbené. Oblíbené stránky jsou ty, na které se hojně odkazují stránky jiné. A konečně u shody lze nastavit stránky přibližně shodné a

přesně shodné. Porovnávána je míra shody slov v dotazu s nalezenými výsledky. Spatřuji zde tedy určitou podobnost s operátorem NEAR. Míra shody jako jediná snižuje důraz na zbylé dva ukazatele. Jezdci určí hodnotu v rozmezí od 0 do 100. Standardně je u všech nastavena hodnota 50. Řazení lze též určit přímo jako součást dotazu. Syntaxe je u všech 3 možností podobná. U aktuálnosti je následující: $\{frsh=hodnota\}$. U oblíbenosti je místo „frsh“ operátor „popl“ a u míry shody „mtch“. Těchto nástrojů lze užít pouze u vyhledávání v sekci „Web“.

Ostatní filtry jsou vesměs podobné s Googlem. Navíc jsou např. filtry „feed“, „contains“ a „url“. „Feed“ podobně jako u Exaleadu dokáže vyhledat RSS zdroje obsahující uvedený výraz. „Contain“ zase vyhledává stránky, obsahující odkaz na uvedený typ souboru. Např. dotaz *contain:mp3 jazz* vyhledá stránky s alespoň jedním odkazem na nějakou MP3ku a slovem „jazz“ kdekoli v dokumentu. Operátor „url“ zjistí, zda je uvedená stránka indexována v databázi Live Search.

4.3.3 Shrnutí

Výhody Live Search spočívají ve velké (kterou ostatně přebíral od MSN) a často aktualizované databázi stránek, plné podpoře booleovských operátorů a ukládáním a zobrazitelném archivu stránek. Možnosti řazení výsledků dle 3 ukazatelů jsou mezi vyhledávacími ojedinělé. Rovněž vzhled vyhledávače je moderní, jednoduchý a pohledný. Nevýhodami jsou absence možnosti zkracování slov, jakýchkoli zástupných symbolů a pokročilého vyhledávání ještě před položením dotazu.

4.4 Jyxo

4.4.1 Obecné údaje

Jyxo je tuzemský fulltextový vyhledávač, specializující se na vyhledávání v českém prostředí. Jedná se o uplatnění technologie Jyxo pro zpracování, analýzu a vyhledávání rozsáhlého množství dat [9]. Tuto technologii využívají též portály Atlas a Quick. Jyxo je modulární vyhledávač, což znamená, že přidání modulů lze využít k prohledávání databází a jiných formátů. V oblibě mají Jyxo hlavně zkušenější a náročnější uživatelé. Index Jyxa obsahuje 97 143 143 dokumentů (stav k 20. 12. 2006) a aktuální údaj o tomto počtu je uveden hned na úvodní stránce vyhledávače spolu s hlavními přednostmi Jyxa – denní aktualizací, českým skloňováním a relevancí výsledků. Vyhledávač je k dispozici na adrese jyxo.cz.

4.4.2 Konkrétní vlastnosti

Úvodní strana je velice úsporná (viz obrázek 17 na další straně). Obsahuje pouze to, co by obsahovat měla. Nic vás na ani neohromí, ani nepohorší. Pole pro dotaz je umístěno v zeleném rámečku přesně ve prostředku stránky. Dotaz se odesílá tlačítkem „Hledat“ či „1.“, jehož význam je stejný jako „I’m Feeling Lucky“ („Zkusím štěstí“) v Googlu. Nad dotazovým polem se nachází možné oblasti hledání. Jsou velmi bohaté. Hledat lze ve stránkách, internetovém zboží, obrázcích, článcích, video a audio souborech a nabídkách dovolených. U všech těchto sekcí hledá Jyxo pouze v České republice, takže se výsledky budou logicky lišit od vyhledávání ve stejných sekcích zahraničních vyhledávačů. To ovšem považuji za skvělý poznatek. Pokud například hledáme nějaký celosvětově známý pojem v obrázcích, pak je pravděpodobné, že budou nalezeny obrázky vztahující se nějakým způsobem k naší vlasti. Stejně je to i u video a audio souborů. Tyto soubory jsou pro nás často zajímavější a tudíž relevantnější. Navíc v databázi obrázků zahraničního vyhledávače by se vyskytovat mohly, ale také nemusely. K vyhledávání v zahraničí slouží karta „ve světě“. Použitý vyhledávač si můžete zvolit v nastaveních. Pod polem pro dotaz jsou odkazy „Nastavení“, „Pokročilé vyhledávání“ a „Přidat stránku“. Význam prvních dvou je zřejmý a třetí odkaz slouží k přidání libovolné stránky psané v češtině do vyhledávací databáze. Můžete přidat až

10 stránek. Nastavení jsou o něco chudší než u zahraničních vyhledávačů. Lze nastavit počet odkazů na stránku, způsob otevírání nalezených odkazů (v novém okně či ne), zapnutí či vypnutí ohýbání slov a doplňování diakritiky a volbu zahraničního vyhledávače pro vyhledávání ve světě. Popis prvků úvodní strany dokončím zmínkou o odkazech na webový log Blog.cz, seznamy článků, moje články, návod na přidání Toolbaru, databázi Usenetu a internetový rozcestník znalostí Mozek.cz.



97 143 143 dokumentů, denní aktualizace, české skloňování, relevance
[Nápověda](#) - [Informace](#) - [Služby](#) - [Kontakt](#) - (C) Jyxo s.r.o.

obr. 17- Úvodní strana vyhledávače Jyxo

A jsme u nástrojů Jyxa. Standardním operátorem je opět AND. Z dalších booleovských operátorů je podporován pouze OR. NOT lze nahradit symbolem „-“. Operátor „+“ má obvyklé použití. Navíc je symbol „#“, který vyjadřuje nepovinnost výskytu slova (tedy obdoba operátoru OPT u Exaleadu). Z proximitních operátorů se připouští pouze fráze. Nejběžnější filtry jsou podporovány. Operátory „inurl“, „intitle“ a „filetype“ však mají v Jyxo podobu „url“, resp. „title“, resp. „format“. Obdobou „language“ je zase „lang“. Syntaxe ostatních operátorů je stejná.

Jyxo má na české poměry několik unikátních vlastností [10]. Umí skloňovat česká slova, což nedokáže ani žádný zahraniční vyhledávač. Na dotaz *vyhledání informací* tak dostaneme dokumenty s výskytem „vyhledáním informace“, „vyhledání informace“ apod. Navíc Jyxo uvažované skloňované tvary i vypisuje. Po zadání několika dotazů mě udivilo, že Jyxo dokázal vyskloňovat také jména. Skloňování slov lze samozřejmě vypnout. Svou databázi aktualizuje Jyxo každý den. Výsledky vyhledávání člení do skupin podle typu příslušnosti. Kliknutím na název nějaké skupiny se dá oblast vyhledávání zúžit. Podobně funguje také Exalead. Na dotaz *vyhledání informací* se v pravé části obrazovky s výsledky objevily skupiny „ČVUT“, „Universita Karlova“ aj. Hodnocení zvnějšku je další vlastností, kterou se Jyxo chlubí. Podstata spočívá v ohodnocení relevantních stránek nejen podle klíčových slov, ale také podle názorů jiných lidí. Je to tedy jakási obdoba PageRanku u Googlu. Jyxo také kontroluje pravopisné chyby a pokud se vyskytnou, pak sám nabízí pravděpodobný správný tvar. Nutno podotknout, že podobnou funkci má hodně vyhledávačů.

Stránka s výsledky je o něco málo odlišná od jiných vyhledávačů (viz obrázek 18). Pole pro dotaz zůstává stejné jako výběr sekcí, ve kterých se má daný dotaz prohledávat. Čas vyhledávání a počet nalezených výsledků jsou samozřejmostí. Pod nimi se však nachází tabulka, obsahující sponzorované odkazy, které souvisejí se zadaným dotazem. Většinou se jedná o nalezené zboží. Záznamy o hitech jsou následující – název stránky, několik prvních slov, URL adresa, odkaz na náhled stránky a další nalezené záznamy ze stejného webu (standardně se vypisují jen 2 hity z každého webu). Zejména náhled stránky je svérázný. Obsahuje totiž pouze textovou strukturu stránky s popisy odkazů. Z náhledu tedy paradoxně nezjistíte, jak stránka skutečně vypadá. Náhledy ve 3 předešlých vyhledávačích byly rozhodně smysluplnější. V pravé části výsledkové stránky se nachází možnost zúžení původního dotazu do některé z nabízených skupin. Pod ní jsou vypsané vyskočované tvary všech slov dotazu, které byly uvažovány při vyhledávání. Skloňování, doplňování diakritiky a odstraňování duplicit, které jsou standardně zapnuty, lze v této části obrazovky vypnout.

The screenshot shows the Jyxo search engine interface. At the top, there's a navigation bar with categories like 'stránky', 'zboží', 'obrázky', 'články', 'video, hudba', 'dovolená', and 've světě'. The search bar contains the text 'Česká spořitelna' and shows 'Hledat' and '1.' buttons. Below the search bar, it states 'Jyxo našlo 126413 dokumentů za 0.05 sekundy'. The main results area displays a list of search results, including 'Banky - firmy na inform.cz', 'ČSOB - banka pro bohatší život', and 'ECM Prague Open 2006 by Česká Spořitelna'. Each result includes a snippet of text and a link to a preview ('Náhled'). On the right side, there's a sidebar with sections: 'Skupiny' (Groups) listing various finance-related sites, 'Volby' (Choices) with options to toggle features like 'Skloňuji a časuji' and 'Doplňuji diakritiku', and 'Ohýbání' (Bending) with options to toggle 'Ohýbám česká podle slovníku' and 'Skloňuji spořitelna'.

obr. 18 - Stránka s výsledky vyhledávače Jyxo

Pokročilé vyhledávání neobsahuje nic nového. Upřesnit lze nutné, nepovinné či nežádoucí výrazy, fráze, umístění klíčových slov (v titulku či adrese), typ souboru, uvažovanou doménu a vypnutí či zapnutí ohýbání, diakritiky a duplicit.

4.4.3 Shrnutí

Jyxo je špička ve vyhledávání v doménách .cz. Ohýbání slov a opravy pravopisných chyb jsou jeho hlavními přednostmi. Dalšími jsou častá aktualizace, velká relevance výsledků hledání (která je však často zpochybňována) a bohatý rozsah oblastí hledání. Skvělá je především databáze video a audio souborů, která obsahuje již více než 2 miliony souborů. Tvrzení o časté aktualizaci není žádným chvástáním. Jyxo opravdu nachází i články staré jeden den. Slabšími stránkami jsou menší databáze stránek, pro mne nepříliš pohledný design a poměrně chudé metody rozšířeného hledání.

4.5 Morfeo

4.5.1 Obecné údaje

Morfeo je dalším z českých fulltextových vyhledávačů. Podobně jako Jyxo se specializuje na vyhledávání v češtině, což je také hlavním mottem jeho úvodní stránky. Je provozovaný společností NetCentrum na portálu Centrum.cz [11]. Morfeo je založen na Open Source technologii Sherlock Holmes, kterou využívá mimo jiné portál Centrum. Velikost databáze indexovaných stránek Morfea je 140 017 683 (k 21. 12. 2006) a Morfeo ji zaktualizuje v rozmezí 3 dnů až měsíce. K dispozici je na adrese morfeo.centrum.cz.

4.5.2 Konkrétní vlastnosti

Úvodní obrazovka je velmi účelně rozvržena (viz obrázek 19). Horní část stránky zabírá vcelku pohledné logo. Pole s dotazem leží zhruba ve 2/3 výšky obrazovky ve světle modrém rámečku. Dotaz se odesílá tlačítkem „Hledej“ napravo od dotazového pole. Nad ním leží nabízené oblasti vyhledávání, které jsou oproti Jyxu o něco chudší. Vyhledávat lze ve stránkách, obrázcích, zprávách a internetovém zboží. Nalezené zprávy a zboží se však již nacházejí na portálu Centrum. Pod rámečkem s dotazem se nachází odkazy na nápovědu a informace pro webmastery a ještě pod nimi odkazy na partnerské weby (např. Centrum, E-Mail, Xchat apod.).

obr. 19 - Úvodní stránka Morfea

Na začátek zmíním ojedinělou technologii Morfea k výpočtu relevance hitů vzhledem k dotazu. Tou je tzv. číslo Q (quality factor) [12]. Podle něho vyhledávač standardně řadí výsledky od nejlepších k nejhorším. Q se počítá na body a je sumou relevancí všech slov a frází uvedených v dotazu, zvýšenou o ohodnocení stránky a dalších až 3 000 bodů v závislosti na jiných faktorech (blízkost slov aj.). Relevance slova je součtem ohodnocení slova (standardně 10 000 bodů, pokud v dotazu neurčíme jinak) a bodů za umístění slova v textu (tj. zda se vyskytuje v nadpise nebo jako klíčové slovo apod., maximum je 2 000 bodů).

Stránka s výsledky na mě působí poněkud rozpačitým dojmem (viz obrázek 20 na následující straně). Každý záznam je totiž roztažen na celou šíři stránky. Informace u nalezených hitů jsou

obvyklé. Navíc je zde hodnota Q a odkaz „details...“, sloužící k zobrazení podrobnějších informací o nalezené stránce. Stejně jako u Jyxa se zde vyskytuje odkaz na další stránky ze stejné domény. Napravo od výsledků se vyskytují sponzorské odkazy, související s dotazem (pokud takové existují).

V nastaveních lze zapnout či vypnout ohýbání slov, kontrolu překlepů, doplňování diakritiky a vyhledávání synonym. Je také možno obohatit informace o nalezeném hitu o katalogové popisky (je-li stránka registrována v katalogu Centrum) a details o všech hledaných slovech. Vypovídací schopnost detailů mi však uniká. Uvedeno je jen jakési číslo a nikde jsem nenašel, co znamená. Obvyklá nastavení zahrnují počet výsledků na stránku, otevírání výsledků v novém okně a maximální počet odkazů z jednoho umístění.

The screenshot shows the Morfeo search engine interface. At the top, there's a search bar with the text 'vyhledávání informací' and a 'Hledej' button. To the right of the search bar are links for 'rozšířené hledání stránek' and 'nastavení', and a link 'návod a tipy'. Below the search bar, there are checkboxes for 'ohýbání slov', 'překlepy', and 'synonyma'. A navigation bar below that has tabs for 'Stránky', 'Obrázky', 'Zprávy', and 'Zboží'. Below the navigation bar, it says 'Na dotaz "vyhledávání informací" našlo Morfeo 6 305 929 odkazů.' and 'Vybírejte si svá synonyma:'. Below this, there's a section 'Odkazy z fulltextu:' followed by a list of search results. The first result is 'Návod k vyhledávání informací v hlasovacích protokolech' with a link to 'http://www.senat.cz/dokumenty/napoveda_hlasovani.html'. The second result is 'Vyhledávání informací | UNICORE nářadí s.r.o.' with a link to 'http://www.unicore.cz/katalog/vyhledavani/pokrocile.asp'. The third result is 'Vyhledávání informací | PRINT SMART S.R.O. - DEMOVERZE' with a link to 'http://www.printsmartshop.cz/katalog/vyhledavani/pokrocile.asp'. The fourth result is 'Vyhledávání informací | VRABEC A VRABEC s.r.o.' with a link to 'http://shop.vrabecavrabec.cz/katalog/vyhledavani/pokrocile.asp'. The fifth result is 'Webový portál věnovaný syntaktické analýze přirozeného jazyka' with a link to 'http://www.tl.muni.cz/~xpavlov/xml/lexamples/bc2/bc2.html'. The sixth result is 'vyhledávání informací' with a link to 'http://vydavatelstvi.vschl.cz/knihy/uid_es-005hesla/vyhledavani_informaci.html'.

obr. 20 - Stránka s výsledky Morfea

Standardním operátorem je i u Morfea AND. Syntaxe a užití operátorů se liší v závislosti na tom, zda se jedná o jednoduché či pokročilé vyhledávání. To je něco, s čím jsem se dosud nikde nesetkal. Nejprve uvedu pravidla jednoduchého vyhledávání. Náhražkou booleovského operátoru OR je zde symbol „?“ , místo NOT lze užít běžného „-“. Fráze mají obvyklé použití. Morfeo má obdobu proximitního operátoru NEAR, je jím symbol „*“ . Nelze však určit, jak daleko od sebe mají být slova vzdálena. Standardně se hledají co nejbližší u sebe. Je třeba dávat pozor, aby mezi slovy a symbolem byla mezera a celý výraz byl v uvozovkách. Dotaz „thomas * edison“ najde jak výraz „thomas edison“, tak výraz „thomas alva edison“. Zajímavé je, že u Morfea funguje zástupný symbol „*“ , který slouží ke zkracování slov. Před ním je však potřeba zadat minimálně 3 písmena a celý výraz nesmí být v uvozovkách. Tento znak má tedy v Morfeu dvojí použití. Pokud se dá mezi slova (oddělí se od nich mezerou a celý výraz se dá do uvozovek), funguje jako NEAR. Uvedeme-li ho za kořen slova (a bez uvozovek), funguje jako prostředek k jeho zkrácení. Nejdůležitější filtry jednoduchého vyhledávání jsou „title“ (vyhledávání v titulku), „hdr“ (vyhledávání v nadpisu úrovně 1 až 6), „keyword“ (slovo se musí vyskytovat jako klíčové) a obvyklý filtr „link“.

U pokročilého vyhledávání je syntaxe o něco složitější. Že se jedná o pokročilý dotaz, sdělím vyhledávací symbolem „!“ (vykřičník). Pokročilé vyhledávání se od jednoduchého liší v tom, že umožňuje výskyty slov kombinovat pomocí booleovských operátorů a ovlivňovat třídění výsledků. Základním pravidlem pokročilého vyhledávání je, že se i klíčová slova píšou v uvozovkách. Obvyčejné slovo zde od fráze tedy neodlišíme, resp. odlišíme tím, že fráze má obvykle více slov.

Operátory AND, OR, AND NOT a užití závorek mají obvyklé použití. Symbol „*“ má použití jako v jednoduchém vyhledávání. Pokud navíc klíčová slova (ta v uvozovkách) oddělíme tečkou, pak lze užít též operátorů NOT a MAYBE. Význam NOT je zřejmý a MAYBE vyjadřuje nepovinnost výskytu slova.

V pokročilém vyhledávání lze rovněž zvýšit ohodnocení slova z implicitních 10 000, což považují za velmi užitečný nástroj. Syntaxe je jednoduchá. Dotaz ! „vyhledávání“/30 000 AND „informací“ přiděluje slovu „vyhledávání“ větší důležitost. Rozsah možného ohodnocení se pohybuje v rozmezí od -30 000 do +30 000. Filtry pokročilého vyhledávání jsou vesměs stejné jako u jednoduchého vyhledávání. Důležitými filtry navíc jsou „URLWORD“ (vyhledání slov v URL adrese stránky, tedy obdoba „inurl“), „EXT“ (slova v odkazech na stránku) a obvyklé „FILETYPE“, „DOMAIN“ a „LANG“. Rovněž syntaxe je poněkud odlišná. Operátory se musí psát velkými písmeny a jejich hodnoty musí být v uvozovkách. Např. požadavku vyhledání slova „informace“ v URL adrese by odpovídal dotaz ! URLWORD „informace“. Zajímavý je operátor „AGE“, sloužící k vyhledání stránek o určitém stáří. Dotaz ! AGE < 604800 „informace“ vyhledá stránky se slovem „informace“, které jsou novější než 7 dní (hodnota musí být uvedena v sekundách). Jako srovnávacích operátorů lze užít také >, =, <=, => a <>.

Za velmi užitečné považují zkratky Morfea, které slouží k vyhledávání specifických informací. Zmíním zkratky „mapy“ (vyhledávání v mapách), „zboží“ (vyhledávání v internetovém zboží), „slov“ (překládání slov), „zprávy“ (vyhledání ve zprávách, nad a je skutečně čárka) a „obr“ (vyhledávání obrázků). Syntaxe je ve formě *nazev_znacky: hodnota*. Např. dotaz *zprávy: topolánek* vyhledá v sekci zpráv články s výskytem slova „topolánek“. U použití slovníku je nutné znát zkratky jazyků. Písmeno „c“ označuje češtinu, „a“ angličtinu, „n“ němčinu, „f“ francouzštinu, „i“ italštinu a „r“ ruštinu. Dotaz *slov ca: pes* přeloží slovo pes z češtiny do angličtiny. Pro otočení směru překladu stačí prohodit písmena.

Ostatní vlastnosti zahrnují podobně jako u Jyxa skloňování a časování slov, automatické opravy překlepů a gramatických chyb a nabízení synonym k danému dotazu. To vše lze samozřejmě také vypnout. Pokročilé možnosti vyhledávání jsou totožné se zmíněnými operátory. K dispozici je vyhledávání v 5 typech souborů.

4.5.3 Shrnutí

Výhodami Morfea jsou poměrně slušná databáze stránek (avšak menší rozsah oblastí), možnosti ohýbání slov, nabízení synonym k vyhledávaným výrazům, opravy překlepů a vcelku příjemný vzhled. Za velmi výhodné považují možnosti využití zkratk a nacházení stránek o určitém stáří. Jako nevýhody musím uvést velmi kostrbaté používání filtrů a dalších pokročilých operátorů. Zdá se mi, že tvůrci z jednoduchých věcí učinily věci složité jen proto, aby se jejich vyhledávač něčím odlišoval.

5 Případová studie

5.1 Úvod

Studie má za cíl srovnat chování různých typů uživatelů při vyhledávání informací v Internetu. Zvolená metoda vychází ze studií, se kterými jsem se setkal v průběhu této práce. Zejména se jedná o studii Anne Auly z univerzity v Tampere, která však byla zaměřena speciálně na formulaci dotazu uživatelů. Studie firmy iProspect zase zkoumala chování uživatelů při vyhledávání obecně. Následující studie se nachází někde mezi nimi. Na jedné straně jsem dotazovaným ponechal možnost výběru při volbě způsobu vyhledání. Nepožadoval jsem tedy nutně vyhledání pomocí vyhledávacích strojů. Pokud se však dotazovaný rozhodl pro nějaký vyhledávací stroj, položené dotazy jsem zanalyzoval.

Jelikož je tato studie jen součástí většího celku, množství úloh a počet dotazovaných je logicky menší než u kvalifikovaných a pro profesionální účely prováděných studií. Pro průzkum jsem vybral 6 otázek, které jsem položil 6 uživatelům s vysokoškolským vzděláním (u třech lidí s dokončeným, všichni na VŠE v Praze). Uživatele jsem vybíral tak, aby svými zkušenostmi s vyhledáváním a počítači obecně reprezentovali 3 kategorie nejčastějších uživatelů. První kategorie zahrnuje méně zkušené, druhá středně pokročilé a třetí Internetu znalé uživatele.

Všech šest otázek jsem zvolil poměrně konkrétně. Obecná témata zavdávají příčiny k nedorozuměním. Nemuselo by totiž být zřejmé, zda uživatel našel přesně to, co jsem po něm požadoval. Konkrétní otázky vylučují spekulace, protože je k nim jasná a jednoznačná odpověď. Pro výzkum jsem zvolil **následujících 6 otázek**:

- [1] V lednu 2007 chci jít do pražského kina Aero na film Plechový bubínek. Kdy ho tam dávají?
- [2] Nalezněte 2 filmy z roku 1940 s Josefem Kemrem coby hercem?
- [3] Kolik členů má NATO a je jeho členem také Turecko?
- [4] Co znamená slovo indigenace?
- [5] Kolik litrů krve proběhne srdcem za 1 minutu?
- [6] Jaký zní španělský název strany venezuelského prezidenta Huga Chaveze, jakou má strana španělskou zkratku a jakou barvu má její logo?

Správné odpovědi zní:

- [1] V sobotu 13. ledna 2007 od 17.45.
- [2] např. To byl český muzikant a Prosím, pane profesore.
- [3] 26, ano.
- [4] Proces uzpůsobování věcí tak, aby vyhovovaly místní (domácí) kultuře.
- [5] 5 – 6 litrů za minutu
- [6] Movimiento V (Quinta) República, MVR, logo má červenou či žluto-bílou barvu

5.2 Průběh studie

Proces vyhledávání odpovědí probíhal následujícím způsobem. Pořadí dotazovaných je uvedeno od uživatele nejméně zručného po uživatele nejzkušenějšího. V hranatých závorkách je uvedeno číslo otázky.

Respondent č. 1 – 49 let, žena, malé zkušenosti s vyhledáváním, počítač používá hlavně k psaní emailů a čtení zpravodajských novinek:

- [1] Uživatelka nejdříve zvolila velmi netradiční přístup, metodu hádání URL adresy. První pokus byl www.kino.cz, což nevede zcela do ztracena. Na daném umístění se totiž vyskytuje odkaz na programy kin ČR (dokina.tiscali.cz). Tato stránka již obsahuje

- vyhledávací formulář, který však po zadání města a názvu filmu nevedl k nalezení požadovaného představení. Server totiž obsahuje jen programy kin na stávající měsíc. Druhý pokus o uhádnutí (oficiálních stránek kina) byl již úspěšný – www.kinoaero.cz. Zde se nachází odkaz na měsíční program, který však znázorňuje měsíc prosinec. Je pravda, že odkaz na následující měsíc (zde leden) je na stránce velmi nenápadný. Respondentka si ho bohužel nevšimla a tento úkol vzdala, tak jsem jí ho raději ukázal. Poté již bylo velmi snadné nalézt v programu kýžený film.
- [2] Uživatelka zkusila nejdříve stejnou metodu jako v předchozím případě. Odhad zněl www.archivfilmu.cz. Taková stránka však neexistuje. Další odhad byl www.filmy.cz. Je zajímavé, že tato URL vede na stránky vyhledávače Centrum. Uživatelka zde zkusila hledání v katalogu, poprvé v adresáři „Vzdělání“ a následně v adresáři „Zábava“. Samozřejmě neúspěšně. Poté položila poměrně slušný dotaz *Kemr film 1940*. Prvním smysluplným hitem byl ten třetí, který také obsahoval potřebnou filmografii.
- [3] Především úkol uživatelku přesvědčil, že bude lepší vyhledávat pokládáním dotazů. Zkusila tedy štěstí dotazem *NATO*. Uspokojivý se zdál být 3. hit – oficiální portál Informačního centra o NATO. Zde respondentka intuitivně klikla na odkaz „Informační centrum o NATO“. Na následující stránce zkusila nejprve odkaz „Historie“, po nenalezení údajů následoval odkaz „Důležité dokumenty“. Na tomto umístění se již nachází odkaz „Členské státy NATO“, obsahující požadované informace.
- [4] První dotaz byl vcelku logicky pojem *indigenace*. Vyhledávač našel pouze 1 odkaz, který odkazoval na jakýsi dokument .pdf, shodou okolností umístěný v doméně vse.cz. Je zajímavé, že tento dokument sice obsahuje odpověď na žádanou otázku, avšak vyhledávač uvedl ve výsledcích část textu souboru, která pojem vůbec neobsahovala. Uživatelce tak přišel hit logicky irelevantní. Dalším krokem byl opět odhad URL adresy a opět ne zcela scestný – www.slovníkicizichslov.cz. Odkaz vede na on-line slovník cizích slov. Pojem *indigenace* zde však nenaleznete. V pravém horním rohu se však nachází logo Googlu včetně textového pole pro vyhledávání. To uživatelku upoutalo a zadala pojem vyhledat Googlu. Druhý nalezený hit (celkem 3 hity) obsahoval odpověď. Byl to však stejný soubor jako ten nalezený vyhledávačem Centrum. Avšak text zobrazený ve výsledcích tentokrát obsahoval část s odpovědí.
- [5] Respondentka opět zkusila odhady umístění. Avšak ani jedna z adres www.lidsketelo.cz, www.lidskeorgany.cz a www.medicina.cz nevedla k úspěchu (první dvě umístění vůbec neexistují). Poté přišly na řadu dotazy. Nejprve *srdce*, pak *lidské srdce*, následně *lidské tělo* a nakonec *lidské orgány*. Stránka s výsledky tohoto dotazu obsahovala odkaz do katalogu firem – lidské orgány, avšak ani tento krok nikam nevedl. Následoval poslední dotaz *наука о lidských orgánech*. Po prozkoumání několika hitů uživatelka snažení vzdala.
- [6] Prvotní dotaz zněl *Venezuela*, což vedlo na cca 200 tisíc odkazů. Další dotaz byl již konkrétnější – *Hugo Chavez*. Uživatelka zkusila hned 1. odkaz na českou verzi encyklopedie Wikipedia, avšak český článek požadované informace neobsahoval. Po tomto neúspěchu dotazovaná své snažení opět vzdala.

Respondentka neměla zkušenosti s vyhledáváním, a tak často volila odhad URL adresy. Na stránky vyhledávače Centrum se dostala zcela náhodně (také odhadem URL). Z toho také vyplývá, že jí bylo lhostejno, jaký vyhledávač používá. Centrum užívala po celý zbytek vyhledávání. Stejně tak uživatelka nevěděla, že vyhledávač nerozlišuje malá a velká písmena. Patrná byla naprostá nezkušenost v pokládání dotazů. Nejenže byly často obecné, ale i věcně nesouvisící s požadovaným tématem.

Respondent č. 2 – 49 let, muž, vyhledává poměrně často, avšak jen na serveru Seznam, kde má také emailovou schránku, Internet užívá k přímému bankovníctví, psaní emailů, online nákupům:

- [1] První úvahy směřovaly k užití katalogu Seznam. Uživatel zkusil adresář „Kina“, který přímo obsahuje programy kin, ale jen na tento měsíc. Dalším krokem byl dotaz *kino Aero*, u kterého uživatel využil automatického doplňování dotazu Seznamem a přetvořil původně zamýšlený dotaz na *kino aero, biskupcova 31, praha 3-program*. Z výsledků hledání vybral dotazovaný odkaz na oficiální stránky kina Aera, který se však díky snaze o nejpresnější vyjádření nacházel až na druhé stránce s výsledky. Pokud by byl zadán původně zamýšlený dotaz *kino Aero*, byl by odkaz hned na 1. místě. Postup nalezení filmu v programu kina byl již shodný s předchozí respondentkou. Avšak s tím rozdílem, že si uživatel odkazu na lednový program všiml sám.
- [2] První a jediný dotaz na Seznam zněl vcelku logicky *Josef Kemr*. Uživatel použil hned 3. odkazu na životopis tohoto herce v české encyklopedii Wikipedia, který obsahuje též hercovu filmografii.
- [3] Prvotním dotazem byl pojem *NATO*, u něhož uživatel opět využil automatického doplňování dotazu, který tak nakonec zněl *nato a členské státy*. Poměrně důvěryhodně vypadal 3. hit „Informační centrum o NATO“, umístěný na serveru www.nato.cz. V sekci „Co je NATO? -> Členské země“ našel uživatel seznam zemí, který obsahoval rovněž Turecko. Testovaná osoba tedy spočetla země, které seznam obsahoval. Na základě tohoto součtu vyřkla nesprávnou odpověď – číslo 19. A to přestože je pod seznamem výslovně uvedeno, že se jedná o údaj z roku 2002.
- [4] Primární dotaz zněl *slovník cizích slov*. Respondent vyzkoušel zadat slovo do prvních dvou slovníků, avšak neúspěšně. Druhý slovník byl stejný jako ten uhádnutý u respondentky č. 1. Rovněž tohoto uživatele napadlo zadat termín do textového pole Googlu v pravém horním rohu. Rovněž on uspěl.
- [5] Uživatel zvolil povedený dotaz *srdce encyklopedie*. Hned 1. odkaz vede na článek o srdci z encyklopedie Wikipedia. Uživatel ho však celý pročetl, než se dostal ke kýžené informaci.
- [6] První dotaz na Seznam zněl zcela nelogicky *Venezuela*, čímž se dotazovaný odsoudil k procházení velkého množství turistických informací. Další dotaz byl již lepší – *Hugo Chavez*. Ten přinesl o prezidentovi množství článků, avšak jméno jeho strany se v několika málo článcích, které uživatel prohlédl, nevyskytovalo. A už vůbec ne ve španělštině. Další dotazy zněly *politický systém Venezuela, politické strany Venezuela* a konečně *strana Hugo Chavez*. Žádný z nich však nevedl k získání požadovaných informací. Nakonec uživatel využil encyklopedie Wikipedia, která mu posloužila v minulém úkolu. Našel článek o Chavezovi v češtině, kde se informace také nevyskytuje. Nakonec jsem musel uživateli poradit, aby zvolil jiný jazyk, kterému rozumí a o němž si myslí, že by v něm o hledané osobě mohlo být více informací. V německé verzi článku se již informace nachází. Logo však uživatel nikde nenalezl.

Respondent dopadl rozhodně lépe než ten první. I on však při vyhledávání nadělal několik chyb. Ani on netušil, že velikost písmen vyhledávač nerozlišuje. Ve 3. otázce důvěřoval uživatel údaji z roku 2002 a neprovedl žádné ověření informace, z čehož plynula chybná odpověď. V zájmu zrychlení vyhledávání nepoužil respondent procházení stránky (CTRL + F), takže článek o srdci četl celý dokud nenarazil na požadovanou informaci. V poslední otázce užil nejprve velmi širokého dotazu. Respondent byl při vyhledávání značně netrpělivý. Nalezenými hity procházel velmi rychle a ani jednou se nevzdálil na více než 2 navigační úrovně od vyhledávací stránky Seznamu (což však nemusí být vždy chybou).

Respondent č. 3 – 25 let, žena, počítač využívá každý den, k vyhledávání používá výhradně Googlu:

- [1] Uživatelka znala URL stránek kina Aero z paměti, takže si na oficiálních stránkách našla odkaz na měsíční program a dokonce si všimla nenápadné ikony na měsíc leden.
- [2] Také zde uplatnila uživatelka znalost vyhledávaného oboru. První kroky směřovaly na server České filmové nebe (www.cfn.cz), kde si filmografii Josefa Kemra vyhledala v databázi.
- [3] Prvotní dotaz byl položen jako *ministerstvo obrany*. Na stránkách Ministerstva obrany ČR (www.army.cz) provedla uživatelka vyhledání pojmu *členové NATO* a následně jen *NATO*. Obě hledání byla neúspěšná. Uživatelka si poté všimla odkazu „Legislativa, NATO, EU -> NATO – odkazy“, kde se rovněž vyskytuje odkaz na oficiální portál informačního centra o NATO. Další postup byl již zcela totožný s postupem 1. respondentky po nalezení tohoto portálu (včetně nesprávného odhadu směřujícímu k historii NATO).
- [4] Uživatelka nejprve položila stejný dotaz jako její předchůdce – *slovník cizích slov*. Po třech marných pokusech o nalezení pojmu ve slovníku položila respondentka pojem přímo vyhledávači. A uspěla.
- [5] První dva dotazy – *krevní oběh* a *krevní oběh srdce* – byly pro uživatelku neúspěšné. Neřekl bych však, že to byly špatně položené dotazy. Uživatelce se spíše nechtělo procházet množstvím výsledků. Místo toho ji napadlo zvolit encyklopedii Wikipedia a zadat dotaz *krevní oběh srdce*. Hned první nalezený odkaz se týkal srdce. V něm se uživatelka pročetla ke správné odpovědi.
- [6] Respondentka rovnou zvolila encyklopedii Wikipedia. Nejprve zadala jméno prezidenta české verzi (zde název strany není), poté anglické. V anglické verzi se informace vyskytuje, avšak uživatelka ji nenalezla a v celé stránce se ani nepokusila vyhledat termín „party“ (strana). Její název však našla v německé verzi článku. Poté klikla na název strany, ale logo se v německém článku o ní nevyskytuje (v anglickém ano). Uživatelka tedy zadala španělský název strany do Googlu. Hned první hit odkazoval na oficiální stránky strany (pouze ve španělštině), které však logo neobsahovaly. Úspěšný byl až 6. hit dotazu *Movimiento V República flag*, který požadované logo obsahuje.

Vyhledávání této uživatelky poukazuje na dvě věci. Tou první je skutečnost, že znalost vyhledávaného oboru pomáhá při vyhledávání (viz první 2 otázky). Tou druhou je názorná ukázka postupu od obecného pojmu ke konkrétnějším informacím (zejména u 3. otázky). Ani tato respondentka však netušila, že vyhledávače nerozeznávají velká a malá písmena. Ani ona nevyhledávala uvnitř stránek, a tak článek o srdci četla, dokud nenalezla výsledek. Také o vyhledávání v obrázcích Googlu uživatelka vůbec netušila. Logo strany by s jeho použitím získala o hodně rychleji.

Respondent č. 4 – 27 let, muž, počítač užívá každý den již od mládí, k vyhledávání používá Googlu a někdy také Yahoo!:

- [1] Uživatel nejprve zkusil server dokina.tiscali.cz, který využívá ke zjištění programu kin. Nevěděl však, že programy na následující měsíc zde nejsou. Dalším krokem byl velmi konkrétní dotaz *plechový bubínek aero program leden*. Již 3. odkaz na magazín Houser byl úspěšný.
- [2] Naprosto totožný postup s předchozím uživatelem.
- [3] Uživatel okamžitě zvolil českou verzi encyklopedie Wikipedia a dotaz *nato* byl úspěšný.
- [4] Prvotní volbou byla opět encyklopedie Wikipedia. Tentokrát však bezvýslednou. Dalším krokem byl dotaz *slovník cizích slov* pro Google. Uživatel vyzkoušel jen jeden slovník, avšak marně. Poté zkusil dotazy *indigenace význam* a *indigenace je*. Ani jeden nepovažoval za dost dobrý, přestože oba přinesly správnou odpověď. Uživatel totiž nekliknul ani na jeden z hitů (byť byly jen 3). Nakonec se vrátil k Wikipedii, jejíž anglické verzi položil dotaz *indigenation*. Toto slovo odhadl jako anglický překlad slova

indigenace. Čtvrtý odkaz směřoval na slovo „indigenization“, což je správný překlad indigenace. Odpověď získal přeložením definice.

- [5] Dotazovaný opět zvolil encyklopedii Wikipedia. Zadal jí pojem *krevní oběh*. Z nalezených záznamů nejprve zvolil „krev“, poté „srdce“. V článku o srdci, který obsahuje odpověď, ale taktéž nevyhledával.
- [6] Primárním krokem byla opět Wikipedia. Nejprve zadal do české verze dotaz *hugo chavez*. Jeden ze dvou odkazů ho přivedl na článek „Hlavy států a organizací 2006“. Uživatel klikl u státu Venezuela na odkaz „Venezuela“. Název strany měl být ve španělštině, a tak uživatel přepnul informace o Venezuele do španělštiny. Poté klikl v souhrnném rámečku na pravé straně na odkaz „Hugo Chávez“. Španělský článek o prezidentovi již obsahoval název jeho strany včetně zkratky. Hledání loga však bylo kostrbaté. Uživatel nezkusil anglickou verzi článku o prezidentově straně, ale pokoušel se najít logo v sekci „Web“ vyhledávače Google. Po dotazech *la quinta republica chaves logo* a *la quinta republica logo* přišel již úspěšný dotaz *MVR logo*. Hned 1. hit obsahoval obě verze loga a byl to stejný hit jako u předchozí respondentky.

Rovněž tento uživatel využil při vyhledání prvních dvou odpovědí své znalosti vyhledávané oblasti. Uživatel věděl, že vyhledávače nerozlišují velká a malá písmena. K tíži bych uživateli přičetl, že v otázce 4 měl již odpověď v podstatě vyhledanou, ale nedokázal ji z hitů vyčíst. Podobně tomu bylo u 6. otázky. Neznalost vyhledávání v obrázcích je taktéž zarážející.

Respondent č. 5 – 23 let, žena, studentka 4. fakulty na VŠE v Praze, počítač využívá každý den, vyhledává každodenně, většinou pomocí vyhledávače Seznam:

- [1] Původní dotaz zněl *kino aero*. První hit dovedl uživatelku na oficiální stránky kina. V měsíčním programu si však uživatelka nevšimla odkazu na měsíc leden, což je ale do značné míry chyba tvůrce stránky. Další dotaz zněl *kino aero leden*. Hned 1. hit odkázal uživatelku na kalendář kulturních akcí na leden, kde byl zmíněn i film Plechový bubínek, avšak bez data vysílání. Úspěšný byl až dotaz *kino aero leden plechový bubínek*. Šestý odkaz směřoval na kulturní program s požadovanou odpovědí.
- [2] Hned prvotní dotaz byl úspěšný – *josef kemr 1940*. Osmý odkaz obsahoval filmografii tohoto herce (také na serveru Českého filmového nebe).
- [3] Původní dotaz zněl *členové nato*. Je zajímavé, že uživatelka při pohledu na 3. hit (článek o Dánsku na Wikipedii) přešla pro vyhledání informací o členech NATO k encyklopedii Wikipedia. Tento hit jí tedy pouze připomněl, že by Wikipedii mohla použít. Zde již uživatelka zadala úspěšný dotaz *nato*.
- [4] Počáteční volbou byl dotaz *slovník cizích slov*. Uživatelka vyzkoušela 4 slovníky. Ani v jednom se však pojem nevyskytoval. Poté zvolila přímou cestu dotazem *indigenace*, který již byl úspěšný.
- [5] Hned první dotaz byl úspěšný – *krev srdce*. Šestý odkaz obsahoval odpověď (nejednalo se však o vševědoucí Wikipedii). Na stránce ji uživatelka vyhledala efektivně – CTRL + F a zadáním pojmu „litr“.
- [6] Prvním krokem byl dotaz *hugo chavez*. Pátý odkaz obsahoval stručný životopis, avšak bez španělského názvu jeho strany. Další dotaz zněl *strana hugo chavez*. První (a také poslední zvolený) hit uživatelku neuspokojil. Zvolila tedy opět Wikipedii. Dotaz pro ni zněl *hugo chavez* a následná navigace vypadala takto: Hlavy států a organizací 2006 -> Venezuela -> prezident Hugo Chávez Frías -> anglická verze článku -> Fifth Republic Movement.

Tato uživatelka prokázala při vyhledávání slušný cit pro tvorbu dotazů, který pramení z jejích zkušeností s vyhledáváním. Navíc se nikdy nedostala do úzkých. Vždy měla po ruce další nápad, jak dotaz zlepšit. Bylo evidentní, že dotazovaná přemýšlela u volby, na který hit kliknout („až“ 8. hit ve 2. otázce). Věděla také, že vyhledávače nerozlišují velká a malá písmena a jako první vyhledávala i uvnitř stránky.

Respondent č. 6 – 25 let, muž, student 4. fakulty na VŠE v Praze, počítač užívá každý den, vyhledává rovněž denně, většinou pomocí Googlu

- [1] Hned prvním dotazem *plechový bubínek kino aero* našel uživatel odpověď na 10. hitu (avšak z jiného zdroje než v předchozích případech).
- [2] Uživatel zvolil databázi celosvětových filmů www.imdb.com. Dotaz byl vcelku logický – *josef kemr*. Fungoval úspěšně.
- [3] Vůbec prvním dotazem *nato počet členů* našel bleskurychle odpověď (dokonce jako vůbec 1. hit).
- [4] Pojem indigenace zadal ihned jako dotaz do Googlu. Řešení je to nejkratší a úspěšné.
- [5] Úspěšné vyhledávání pokračovalo dotazem *srdce litrů krve*, který nachází odpověď hned na 1. hitu (dokonce se vyskytuje ve vybraném textu u výsledku vyhledávání).
- [6] Původní dotaz zněl *hugo chavez strana*. Na 1. stránce s výsledky klikl uživatel na 2 odkazy, ale marně. Zajímavý je způsob, jak respondent na odpověď nakonec přišel. Na 5. stránce s výsledky spatřil odkaz do Wikipedie, avšak na článek o jménu Hugo. Odkaz mu nicméně připomněl, že by mohl k vyhledání užít této encyklopedie. Další postup byl již totožný jako u 5. respondentky.

Bezpochyby nejlepší a nejefektivněji vyhledávající uživatel této studie. To se ukázalo zejména v 5. otázce, kdy odhadl, že by odpovídající dokument mohl obsahovat sousloví „litrů krve“ takto pohromadě. Pro názornější a preciznější použití mohl položit dotaz *srdce „litrů krve“*, ale Google se vždy snaží vyhledávat slova v dotazu primárně jako frázi, tudíž to tam být nemusí. Podle mne uživatel odvedl při vyhledávání velmi kvalitní práci. Jako jediný z respondentů byl až na 5. straně s výsledky, která ho v daném případě nasměrovala k úspěšnému vyhledání.

5.3 Shrnutí studie

Myslím, že studie potvrdila obecnou platnost studií rozsáhlejších. Průměrný počet slov na jeden dotaz se skutečně pohyboval okolo 2 – 3 slov. Až na nejzkušenějšího uživatele nepohlédli nikdo dále než na 2. stránku s výsledky. Konkrétnější dotazy byly vždy úspěšnější než dotazy širší. Nikdo z uživatelů neklikl po zadání jednoho dotazu na více než 3 hity. Uživatelé používali dvou nejrozšířenějších vyhledávačů v ČR – Seznamu a Googlu. Znalost vyhledávané oblasti dotazovaným významně pomohla. Operátory nepoužil nikdo.

Ze studie jsem však vypožoroval další zajímavé skutečnosti. Pro vyhledávání v Internetu je velmi důležitá znalost cizího jazyka. Nejlepší je, pokud uživatel ovládá angličtinu. V té se totiž vyskytuje největší množství informací. Pokud se jedná o nějakou obecnou a celosvětově zkoumanou oblast, pak bych vždy položil dotaz v angličtině. Nikdo z dotazovaných při nalézání loga nevyužil vyhledávání v databázi obrázků. Této možnosti velmi hojně využívám. Vždyť databáze obrázků Google obsahuje přes miliardu obrázků. Dalším zajímavým zjištěním pro mne byla hojnost využití internetové encyklopedie Wikipedia. I zde platí, že znalost angličtiny je výhodou. V tomto jazyce se na Wikipedii nachází úctyhodných 1,5 milionů článků (oproti 55 tisícům českých článků). Navíc Wikipedia funguje nejen jako encyklopedie všech oborů, ale také jako slovník cizích slov a výkladový slovník. Zajímavé také je, že se uživatelé často nacházeli již velmi blízko správné odpovědi (hlavně u 6. otázky). Někdy ji v záplavě textu přehlédli, někdy od ní byli vzdáleni na jedno kliknutí. Ve dvou případech zachránilo uživatele logo Googlu v pravém horním rohu.

Často jsem se zamýšlel, zdali je lepší včas pozměnit dotaz nebo spíše projít více stránek s výsledky. Studie totiž ukázala, že i průměrným dotazem se dá leccos najít. Osobně se přikláním k názoru, že pokud si je uživatel jist svým dotazem (a dotaz je již dostatečně konkrétní, např. 4 slova), měl by projít více stránek s výsledky. Opačným extrémem je, že uživatel neklikne ani na jeden hit (ve studii se tato situace stala dvakrát). Otázkou pak je, proč uživatel vůbec daný dotaz pokládal.

Závěr

Z teoretického hlediska měla práce objasnit základní tři problémy, týkající se vyhledávání v Internetu. Vše jsem se snažil obohacovat o praktické příklady a své připomínky k daným tématům. To se týkalo zejména správného pokládání dotazů (a tudíž nalézání co nejrelevantnějších informací) a názorných ukázek použití operátorů a filtrů.

Z praktického hlediska měla práce poukázat na závislost kvality vyhledaných informací na dvou hlavních faktorech – vhodné volbě vyhledávacího systému a povaze uživatelů. První faktor závisí především na velikosti indexové databáze, podporovaných operátorech, množství dalších oblastí k vyhledávání a rozhraní vyhledávače. Velikostí indexu dominuje vyhledávač Google. Pokud tedy uživatel požaduje co nejvíce informací (tj. úplnost), měl by zvolit Google. Také počet oblastí pro vyhledávání je nejrozsáhlejší u Googlu. Hlavně při hledání obrázku či řešení problému bych doporučil jeho použití. Databáze obrázků a diskusních skupin jsou totiž jedinečné. Stejně tak k nalezení specifických typů souborů bych užil Googlu, který je v tomto ohledu „nejštedřejší“. Vyhledávač Exalead doporučuji k vyhledávání multimediálních typů souborů, zejména audio a video souborů. Vhodný je také k nalezení RSS zdrojů. Možnost využití regulárních výrazů u Exaleadu je výtečná. To platí i o příjemném uživatelském rozhraní. Mile mne překvapil vyhledávač Windows Live Search, a to hlavně relevancí výsledků, která mi často přišla dokonce vyšší než u Googlu. Live Search však ztrácí v oblasti pokročilých možností vyhledávání, které nejsou příliš bohaté. Ty jsou nejpestřejší u Googlu, což se týká hlavně množství operátorů. Exalead se v tomto ohledu nachází zhruba vprostřed.

Ze srovnání českých vyhledávačů pro mne vyšel lépe vyhledávač Jyxo. Velikost indexu má sice menší než Morfeo, avšak pokročilé možnosti vyhledávání jsou názornější a jednodušší. Pro Jyxo též hraje větší rozsah vyhledávaných oblastí a pravidelnější aktualizace indexu. U Morfea oceňuji možnost užití vyhledávacích zkratk a také větší relevanci výsledků, ačkoli je relevance jedním z hlavních hesel Jyxa.

Druhý stěžejní faktor, ovlivňující kvalitu vyhledaných informací, je povaha uživatelů. S tímto tématem souvisí především problémy správného pokládání dotazu a výběru vhodných dokumentů z nalezených výsledků. V praxi jsem zjistil, že základem dotazu jsou z 90 % klíčová slova a jen v menší míře znalost operátorů. Ta však může uživateli společně s užitím filtrů přinést vysoce relevantní výsledky. Operátory a filtry bych tedy užil pro zvýšení přesnosti vyhledávání, tj. k získání konkrétních a značně specifických informací.

Studie v 5. kapitole měla za úkol porovnat výzkumy zmíněné v práci s výzkumem vlastním. Ačkoli byl počet dotazů a respondentů mé studie menší ve srovnání s profesionálně prováděnými studiemi, prokázal signifikantní analogii. Dotazy typického uživatele jsou krátké, uživatel prochází jen velmi málo nalezených hitů a znalost vyhledávaného oboru mu pomáhá při vyhledávání. Platnost obecných axiomů byla tedy dokázána. Domnívám se proto, že práce svůj stanovený cíl splnila ve všech výše zmíněných ohledech.

Seznam použitých zdrojů

Kapitola 1

- [1] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 2
- [2] Internetová encyklopedie Wikipedia, anglická verze, výraz „Information“, 14. 11. 2006, <http://en.wikipedia.org/wiki/Information>
- [3] Internetová encyklopedie Wikipedia, anglická verze, výraz „Paper“, 14. 11. 2006, <http://en.wikipedia.org/wiki/Paper>
- [4] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 181
- [5] Lyman, Peter a Hal R. Varian, „How Much Information“, 2003. výsledky výzkumu dostupné na http://www.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf
- [6] Lyman, Peter a Hal R. Varian, „How Much Information“, 2000. výsledky výzkumu dostupné na <http://www.sims.berkeley.edu/how-much-info>
- [7] Lyman, Peter a Hal R. Varian, „How Much Information“, 2003. výsledky výzkumu dostupné na http://www.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf
- [8] tamtéž
- [9] Internetová encyklopedie Wikipedia, anglická verze, výraz „Information overload“, 16. 11. 2006, http://en.wikipedia.org/wiki/Information_overload

Kapitola 2

- [1] Elektronický časopis Ikaros, Kognitivní aspekty procesu vyhledávání informací. ročník 10. číslo 9 (2006). <http://www.ikaros.cz/node/3592>
- [2] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 28
- [3] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 26 – 27
- [4] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 234
- [5] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 259
- [6] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 260
- [7] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 260
- [8] Internetová encyklopedie Wikipedia, anglická verze, výraz „Search engine“, 19. 11. 2006, http://en.wikipedia.org/wiki/Search_engine
- [9] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 185
- [10] Internetová encyklopedie Wikipedia, anglická verze, výraz „Search engine“, 19. 11. 2006, http://en.wikipedia.org/wiki/Search_engine
- [11] Internetová encyklopedie Wikipedia, anglická verze, výraz „Exalead“, 19. 11. 2006, <http://en.wikipedia.org/wiki/Exalead>
- [12] Internetová encyklopedie Wikipedia, anglická verze, výraz „Search engine“, 20. 11. 2006, http://en.wikipedia.org/wiki/Search_engine
- [13] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 238
- [14] Danny Sullivan, New Estimates Puts Web Size At 11,5 Billion Pages and Compares Search Engine Coverage. 17. května 2005. <http://blog.searchenginewatch.com/blog/050517-075657>

- [15] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 224
- [16] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 231

Kapitola 3

- [1] Rejman, Ladislav, Kapesní slovník cizích slov. Státní pedagogické nakladatelství Praha. 1971
- [2] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 215
- [3] Internetová encyklopedie Wikipedia, anglická verze, výraz „Relevance (information retrieval)“, 26. 11. 2006, [http://en.wikipedia.org/wiki/Relevance_\(information_retrieval\)](http://en.wikipedia.org/wiki/Relevance_(information_retrieval))
- [4] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 215
- [5] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 215
- [6] Creighton University, Measuring Search Effectiveness, 27. 11. 2006, <http://www.hsl.creighton.edu/hsl/Searching/Recall-Precision.html>
- [7] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 216
- [8] Creighton University, Measuring Search Effectiveness, 27. 11. 2006, <http://www.hsl.creighton.edu/hsl/Searching/Recall-Precision.html>
- [9] tamtéž
- [10] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 296
- [11] Internetová encyklopedie Wikipedia, česká verze, výraz „PageRank“, 30. 11. 2006, <http://cs.wikipedia.org/wiki/PageRank>
- [12] Jan Kužník, Google bomba odrovnala Grosse, Bushe i Microsoft. 15. 11. 2004. http://technet.idnes.cz/sw_internet.asp?r=sw_internet&c=A041112_5286450_sw_internet
- [13] http://www.iprospect.com/premiumPDFs/WhitePaper_2006_SearchEngineUserBehavior.pdf
- [14] Belkin, N. J., Helping people find what they don't know. Digitální knihovna ACM. 2000. http://portal.acm.org/citation.cfm?id=345143&coll=portal&dl=ACM&CFID=6408739&CF_TOKEN=39633098
- [15] např. studie provedená Anne Aulou z University of Tampere, dostupná na www.cs.uta.fi/~aula/questionnaire.pdf
- [16] takéž výše zmíněná studie, dostupná na www.cs.uta.fi/~aula/questionnaire.pdf
- [17] takéž výše zmíněná studie, dostupná na www.cs.uta.fi/~aula/questionnaire.pdf
- [18] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 41
- [19] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 42
- [20] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 270
- [21] Wesleyan University, Truncation and wildcards – examples, 2. 12. 2006, <http://wesleyan.edu/libr/tut/rst/truncation-h.html>
- [22] CSA Illumina, Help page, Boolean Operators, 3. 12. 2006 http://www.csa.com/help/Search_Tools/boolean_operators.html
- [23] Laura Cohen, Boolean Searching on the Internet, 5. 12. 2006, <http://www.internettutorials.net/boolean.html>
- [24] tamtéž
- [25] tamtéž
- [26] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 45

- [27] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 275
- [28] Creighton University, Measuring Search Effectiveness, 5. 12. 2006, <http://www.hsl.creighton.edu/HSL/searching/ProximityOperators.html>
- [29] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 290
- [30] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 292
- [31] Internetová encyklopedie Wikipedia, anglická verze, výraz „Credibility“, 6. 12. 2006, <http://en.wikipedia.org/wiki/Credibility>
- [32] <http://captology.stanford.edu/pdf/Stanford-MakovskyWebCredStudy2002-prelim.pdf>
- [33] WebCredible, Beyond Web usability: Web credibility. duben 2004. <http://www.webcredible.co.uk/user-friendly-resources/web-credibility/beyond-web-usability.shtml>
- [34] obsah celé studie je k dispozici na adrese <http://www.consumerwebwatch.org/pdfs/stanfordPTL.pdf>
- [35] tamtéž, výrok se vyskytuje na straně 12
- [36] tamtéž
- [37] 10 nejužitečnějších tipů je zmíněno na adrese <http://credibility.stanford.edu/guidelines/index.html>
- [38] Angela Gunn, In Search of Disclosure. 17. dubna 2003. <http://www.consumerwebwatch.org/dynamic/wc-web-ethics-in-search-of-disclosure.cfm>
- [39] Sklenák, V. a kol., Data, informace, znalosti a Internet. 1. vydání. Praha: C. H. Beck 2001. strana 262
- [40] Paul J. Bruemmer, Give Yourself a Little More Reach. 15. srpna 2001. <http://www.clickz.com/showPage.html?page=865811>
- [41] Greg R. Notess, Search Engine Statistics: Freshness Showdown. 17. května 2003. <http://searchengineshowdown.com/statistics/freshness.shtml>
- [42] Greg R. Notess, Search Engine Statistics: Dead Links Report. 21. února 2000. <http://searchengineshowdown.com/statistics/dead.shtml>

Kapitola 4

- [1] server společnosti Zürich Communications, Inc., studie Search Engine Usage Statistics, dostupná na http://seo.zunch.com/search_engine_usage_statistics.htm
- [2] Jan Ambrož, Vyhledávač Google posílil na českém webu, Centrum ztratilo. 20. prosince 2005. <http://www.lupa.cz/clanky/vyhledavac-google-posilil-na-ceskem-webu/>
- [3] Internetová encyklopedie Wikipedia, anglická verze, výraz „Exalead“, 16. 12. 2006, <http://en.wikipedia.org/wiki/Exalead>
- [4] Internetová encyklopedie Wikipedia, česká verze, výraz „RSS“, 18. 12. 2006, <http://cs.wikipedia.org/wiki/RSS>
- [5] Internetová encyklopedie Wikipedia, anglická verze, výraz „Regular expression“, 18. 12. 2006, http://en.wikipedia.org/wiki/Regular_expression
- [6] Internetová encyklopedie Wikipedia, česká verze, výraz „Google“, 18. 12. 2006, <http://cs.wikipedia.org/wiki/Google>
- [7] Johnny Long, Google Hacking. 1. vydání. Brno: Zoner Press 2005. strana 24
- [8] Greg R. Notess, Review of Live Search. 21. října 2006. <http://searchengineshowdown.com/features/live/review.html>
- [9] informace o vyhledávači Jyxo, 18. 12. 2006, <http://jyxo.cz/d/info>
- [10] tamtéž
- [11] Internetová encyklopedie Wikipedia, česká verze, výraz „Morfeo“, 19. 12. 2006, <http://cs.wikipedia.org/wiki/Morfeo>
- [12] nápověda k vyhledávači Morfeo, 20. 12. 2006, <http://morfeo.centrum.cz/index.php?napoveda=1&sec=mor>