Univerzita Karlova v Praze

Filozofická fakulta

Ústav informačních studií a knihovnictví

Studijní program: informační studia a knihovnictví

Studijní obor: informační studia a knihovnictví

BBS Vojtěch Sedláček, DiS.

FENOMÉN GOOGLE

Diplomová práce

Praha 2007

Vedoucí diplomové práce: PhDr. Richard Papík, Ph.D.

Oponent diplomové práce:

Datum obhajoby:

Hodnocení:

Prohlášení:

Prohlašuji, že jsem diplomovou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje.

V Praze, 23. dubna 2007

……..

podpis diplomanta

Identifikační záznam

SEDLÁČEK, Vojtěch. *Fenomén Google [The Google Phenomenon]*. Praha, 2007. 110 s. Diplomová práce. Univerzita Karlova v Praze, Filozofická fakulta, Ústav informačních studií a knihovnictví 2007. Vedoucí diplomové práce PhDr. Richard Papík, Ph.D.

Abstrakt

Předmětem diplomové práce je podrobná analýza internetového vyhledávače Google z pohledu potřeb informačního specialisty. Cílem práce je popis a rozbor možností, které tento vyhledávač nabízí nejen ve smyslu pokročilých funkcí a operátorů.

V úvodu práce jsou popsány vývojové etapy systémů, vyhledávajících informace v prostředí sítě WWW. Následuje rozbor hlavních principů vnitřního fungování vyhledávače Google. Jádro práce tvoří analýzy základních vlastností vyhledávání (implementace booleovské logiky, stemmingu, proximitního vyhledávání), specifik rešeršního procesu a pokročilých operátorů a funkcí. Pozornost je věnována také odlišnému fungování vyhledávání v českém jazykovém rozhraní.

V závěru práce je provedena analýza konkurenčního prostředí a SWAT analýza Googlu.

Klíčová slova

Vyhledávání informací, internetové vyhledávače, informační analýza, informační služby, hodnocení relevance, uživatelské rozhraní, Google.

Obsah

OBSAH **1**

PŘEDMLUVA **4**

1 ÚVOD **7**

1.1 INFORMAČNÍ PRAMENY 7
1.2 VYMEZENÍ POJMŮ 9
1.3 STRUČNÝ PŘEHLED HISTORICKÉHO VÝVOJE VYHLEDÁVAČŮ 11
1.3.1 Archie, předchůdce vyhledávacích nástrojů *11*
1.3.2 Počátky World Wide Webu *11*
1.3.3 Rozvoj sítě WWW *12*
1.3.4 První katalogy *13*
1.3.5 První plnotextový vyhledávač *13*
1.3.6 Altavista *13*
1.3.7 Google *14*

2 MECHANISMUS FUNGOVÁNÍ VYHLEDÁVAČE GOOGLE **15**

2.1 SHROMAŽĎOVÁNÍ DAT 15
2.2 ZPRACOVÁNÍ INFORMACÍ, VÝPOČET PAGERANKU 17
2.3 VYHODNOCOVÁNÍ UŽIVATELSKÝCH DOTAZŮ 18

3 ANALÝZA UŽIVATELSKÉHO ROZHRANÍ **20**

3.1 DOMOVSKÁ STRÁNKA 20
3.2 I'M FEELING LUCKY 22
3.3 ADVANCED SEARCH 24
3.4 PREFERENCES 25
3.5 LANGUAGE TOOLS 26
3.6 ČESKÁ VERZE UŽIVATELSKÉHO ROZHRANÍ 27

4 ANALÝZA ODCHYLEK VE VYHLEDÁVÁNÍ V ZÁVISLOSTI NA JAZYKOVÝCH NASTAVENÍCH **29**

4.1 PROBLEMATIKA 29
4.2 METODIKA TESTOVÁNÍ 30
4.3 FAKTOR DIAKRITIKY 31
4.4 SROVNÁNÍ VARIANT V MODELOVÝCH DOTAZECH 32
4.4.1 Dotaz "páteř" *32*
4.4.2 Dotaz páteř *33*
4.4.3 Dotaz pater *34*
4.4.4 Dotaz "pater" *35*

4.5 SROVNÁNÍ VARIANT PŘI BĚŽNÝCH DOTAZECH ... 36
4.5.1 Mezinárodní termíny ... 36
4.5.2 Mezinárodní termíny s vazbou na české prostředí ... 37
4.6 ZÁVĚR ANALÝZY ... 37
4.7 DALŠÍ ODCHYLKY LOKÁLNÍCH JAZYKOVÝCH VERZÍ ... 38

5 ANALÝZA ZÁKLADNÍCH VLASTNOSTÍ VYHLEDÁVÁNÍ ... 39

5.1 BOOLEOVSKÁ LOGIKA ... 39
5.1.1 Implementace booleovských operátorů v Googlu ... 39
5.1.2 Srovnání s konkurencí ... 40
5.1.3 Nedostatky booleovského modelu pro vyhledávání dokumentů ... 41
5.2 STEMMING, LEMATIZACE, PRAVOSTRANNÉ ROZŠÍŘENÍ ... 43
5.2.1 Stemming ... 43
5.2.2 Srovnání s konkurencí ... 45
5.2.3 Pravostranné rozšíření (truncation) ... 45
5.3 VZDÁLENOSTNÍ OPERÁTORY ... 46
5.3.1 Rozdělení vzdálenostních operátorů ... 46
5.3.2 Operátor fráze ... 48
5.3.3 Zástupné znaky (wildcards) ... 48
5.3.4 Srovnání s konkurencí ... 51

6 SPECIFIKA REŠERŠNÍ ČINNOSTI V PROSTŘEDÍ GOOGLU ... 52

6.1 ŘÍZENÉ SLOVNÍKY ... 52
6.2 VOLNOTEXTOVÉ VYHLEDÁVÁNÍ ... 52
6.3 VYHLEDÁVÁNÍ Z HLEDISKA KOEFICIENTŮ ÚPLNOSTI A PŘESNOSTI ... 54
6.3.1 Koeficient úplnosti R (recall) ... 55
6.3.2 Koeficient přesnosti P (precision) ... 55
6.3.3 Koeficient přesnosti ve volnotextových vyhledávačích ... 56
6.4 DEFINICE A ROLE RELEVANCE ... 57
6.4.1 Důsledky formálně relevantního vyhledávání pro rešeršní činnost ... 58
6.4.2 Jazykové nároky na rešeršní činnost z hlediska formální relevance ... 59
6.4.3 Relevance, úplnost a přesnost ve vztahu k rešeršní strategii ... 60
6.5 VLIV SOFTWAROVÉHO PROSTŘEDÍ NA REŠERŠNÍ ČINNOST ... 62

7 ANALÝZA POKROČILÝCH FUNKCÍ ... 63

7.1 OPERÁTORY PRO MODIFIKACI DOTAZU ... 63
7.1.1 Synonymní operátor ~ (tilda) ... 63
7.1.2 Operátor `site:` ... 64
7.1.3 Operátor `numrange:` ... 64
7.1.4 Operátory `intitle:`, `allintitle:` ... 64
7.1.5 Operátory `inurl:`, `allinurl:` ... 67

7.1.6 Operátory `inanchor:`, `allinanchor:` ... 68
7.1.7 Text příchozích odkazů jako specifický druh sémantických metadat ... 69
7.1.8 Operátory `intext:`, `allintext:` ... 71
7.2 OPERÁTORY JAKO SAMOSTATNÉ FUNKCE ... 71
7.2.1 Vyhledávání definic, operátor `define:` ... 71
7.2.2 Analýza okolí dokumentu, operátor `link:` ... 76
7.2.3 Analýza okolí dokumentu – srovnání s konkurencí ... 77
7.2.4 Analýza okolí dokumentu – nástroje SEO ... 78
7.2.5 Operátor `patent` ... 81
7.2.6 Ostatní operátory ... 82
7.3 BEZOPERÁTOROVÉ FUNKCE ... 83
7.3.1 Začleňování výsledků ze služby Image Search ... 84
7.3.2 Začleňování výsledků z dalších služeb ... 84
7.3.3 Google Calculator ... 85
8 ANALÝZA KONKURENČNÍHO PROSTŘEDÍ ... 87
8.1 VÝVOJ TRHU ... 87
8.2 YAHOO! ... 88
8.3 LIVE SEARCH ... 90
8.4 ASK.COM ... 91
8.5 OSTATNÍ FORMY KONKURENCE ... 93
9 PORTFOLIO SLUŽEB ... 94
9.1 VÝVOJOVÝ TREND SLUŽEB INTERNETOVÝCH VYHLEDÁVAČŮ ... 94
9.2 VYBRANÉ INFORMAČNÍ SLUŽBY Z NABÍDKY GOOGLU ... 95
9.2.1 Google Book Search ... 95
9.2.2 Google Scholar ... 97
9.2.3 Google Patent Search ... 98
10 SWAT ANALÝZA GOOGLU ... 100
ZÁVĚR ... 103
CITOVANÉ ZDROJE ... 104
DALŠÍ POUŽITÉ ZDROJE ... 109

Předmluva

„What information consumes is rather obvious: it consumes the attention of its recipients. Hence, a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it." [1]

Herbert Simon, držitel Nobelovy ceny za ekonomii

„Getting information off the Internet is like taking a drink from a fire hydrant."

Mitchell Kapor, zakladatel Lotus Software

Vyhledávání informací na internetu je stále běžnějším způsobem uspokojování nejrůznějších lidských informačních potřeb. S rostoucím objemem informací, dostupných na WWW prostřednictvím internetových vyhledávačů je však stále obtížnější se v rostoucí množině vyhledaných, často nerelevantních dokumentů orientovat.

Jedno z možných řešení pro nenáročného uživatele představuje spoléhání se na komplexní algoritmy vyhledávačů. Ty při zpracování dotazu a řazení jeho výsledků automaticky určují podle určitého klíče relevanci a důležitost dokumentů, mohou zohledňovat jazyk a geografickou polohu uživatele, odhadovat jeho zájmy na základě předchozích dotazů apod.

Z pohledu informačního specialisty by měl být však tím, kdo řídí rešeršní proces vždy především uživatel, nikoliv automat.

Informační profesionál při své práci používá zpravidla výrazně odlišné, efektivnější rešeršní nástroje a informační zdroje než jakým je Google. Ne vždy je však nutné, vhodné, či možné těchto prostředků využít – například mimo profesi, v běžném životě. To není však důvod, proč by měl pokročilý uživatel tyto základnější formy rešeršních nástrojů využívat také pouze základním způsobem. I v prostředí Googlu může specialista zúročit své dovednosti, pokud si bude vědom všech pokročilých možností, které internetové vyhledávače nabízejí.

Tato práce má za cíl tyto (mnohdy skryté) možnosti popsat, analyzovat, podrobit kritice a zjistit jejich výhody a nevýhody v porovnání s podobnými nástroji.

Toto téma jsem si vybral proto, neboť nejrůznější, nejen informační aspekty prostředí internetu mě dlouhodobě zajímají. Škálu a různorodost informací, dostupných prostřednictvím služeb a zdrojů webové sítě, považuji za fascinující. V neposlední řadě v prostředí internetových informačních služeb také profesně působím.

Tato práce se původně měla věnovat (v rámci zamýšleného vymezení „fenoménu Google“) také jeho dopadu, působení a vlivu na okolní svět. Sem patří například otázka autorských práv, ochrany osobních údajů nebo vývoje vyhledávacích návyků uživatelů. Analýzy Googlu samotného měly tvořit jen část práce, neboť ohledně jeho fungování již existuje nepřeberné množství literatury. Při provádění předběžného výzkumu bylo ale identifikováno hned několik oblastí, které si zasluhují podrobnou analýzu, neboť jsou v běžných zdrojích popsány jen povrchně nebo vůbec (odlišnost pořádání výsledků v českém rozhraní, interpretace diakritiky, přesné fungování některých pokročilých operátorů a funkcí).

Proto bylo rozhodnuto o užším zaměření původně obecněji koncipovaného tématu, celá práce má tak výrazně analytický charakter. Ze stejného důvodu nebyl nakonec do konečné podoby práce zahrnut ani vlastní výzkum uživatelského chování, který probíhá prostřednictvím dotazníku na webové adrese http://vyhledavace.nazory.cz.

V prvních dvou kapitolách práce je stručně shrnut historický vývoj internetových vyhledávacích strojů až ke vzniku Googlu a nastíněn mechanismus jeho vnitřního fungování. Ve třetí kapitole jsou popsány základní prvky uživatelského rozhraní. Poté již následuje samotné jádro diplomové práce. Čtvrtá kapitola se zabývá detailním rozborem vlivu, který na vyhledávání má volba českého jazykového rozhraní. Následují podrobné analýzy základních funkcí (aplikace booleovské logiky, stemming, vzdálenostní vyhledávání) v kapitole páté. Šestá kapitola popisuje specifika rešeršního procesu v prostředí Googlu. V sedmé kapitole jsou podrobeny analýze pokročilé operátory a funkce. Práci uzavírá analýza konkurenčních vyhledávačů (kapitola 8), vybraných služeb (kapitola 9) a závěrečná SWAT analýza (kapitola 10).

Problematika relevantních informačních zdrojů k práci je popsána v kapitole 1.1. Bylo použito především vlastních analýz, poznatků odborníků publikovaných v internetových zdrojích a z malé části také článků z odborných časopisů vyhledaných prostřednictvím databáze Proquest 5000.

Použitá terminologie byla konzultována s informačními zdroji v oboru a s online Terminologickou databází knihovní a informační vědy. Vymezení nejdůležitějších pojmů obsahuje kapitola 1.2.

Zdroje jsou citovány v souladu s normami ISO 690 a ISO 690-2. Citace v textu jsou prováděné pomocí číselných odkazů na seznam v závěru práce. Čísla v horním indexu odkazují na poznámky pod čarou.

Na tomto místě bych chtěl poděkovat také svému vedoucímu diplomové práce, PhDr. Richardu Papíkovi, PhD. za vstřícný přístup, cenné rady a dobře míněné připomínky.

1 Úvod

1.1 Informační prameny

„Researching a search engine algorithm is kind of like atomic research. The search engine algorithm is a black box that you cannot see inside of, so you can only formulate theories about a specific algorithm by testing its behavior. This is why it is so important to know all of the different search engine operators and how they work.“ [2]

Eileen Kowalski

Tato práce se z velké části věnuje podrobnému zkoumání možností vyhledávání pomocí internetových vyhledávacích strojů. Jak je již patrné z uvedeného citátu, tato oblast je velmi specifická z hlediska relevantních informačních pramenů.

Přesné znění algoritmů, na jejichž základě celý systém Googlu funguje, je přísně střeženým obchodním tajemstvím. Google jej chrání nejen před konkurencí, ale i před samotnými uživateli, resp. před autory a správci webových stránek. Znalost všech konkrétních faktorů, majících vliv na posuzování informační hodnoty dokumentu či na určování jeho relevance k určitému dotazu je snadno zneužitelná.

Dokumentace Googlu k možnostem vyhledávání je velmi stručná i z jiného důvodu. Vzhledem k obrovskému množství indexovaných dokumentů a jejich různorodosti Google ani nemůže popsat či zaručit určitou pevně stanovenou funkčnost vyhledávacích nástrojů.

Analýza pokročilých funkcí a operátorů, pomocí nichž lze v Googlu provádět rešeršní proces, je tak značně problematická. Všechny pokusy o přesný popis se musí zakládat jen na hypotézách, testovaných při pokusech různě velkého rozsahu. Navíc závěry těchto pokusů mají časově neurčitou platnost, protože algoritmus vyhledávacích strojů je neustále vyvíjen a v tichosti obměňován.

Z těchto důvodů se tradiční (především tištěné) informační zdroje této dynamické a ne zcela transparentní problematice příliš nevěnují. Odborná periodika v oblasti online informačních služeb (především Online, Searcher, Information Today) se tak zabývají spíše obecnějšími a dlouhodobějšími tématy, jakými jsou nové projekty

informačních služeb, problematika autorských práv, výzkumy uživatelských zvyklostí, budoucnost vyhledávání informací apod. Jednou z mála výjimek je pravidelný sloupek George Notesse „Search Engine Update“ v časopise Online.

Ve sféře čistě internetových, méně formalizovaných informačních zdrojů je situace odlišná. Existuje celý jeden obor, který velmi podrobně sleduje a zaznamenává změny ve fungování internetových vyhledávačů. Je jím optimalizace webových stránek pro vyhledávače (SEO), soubor činností, jejímž cílem je propagace a zviditelnění webových stránek (především z komerčních důvodů).

SEO se zabývá především rozborem algoritmů, kterými vyhledávače posuzují relevanci a citovanost dokumentů. Poznatky z oblasti SEO jsou proto pro rešeršéra využitelné jen částečně. Primárním cílem SEO je vývoj technik, jak informace zviditelňovat, nikoliv analyzovat způsoby jejich pokročilého vyhledávání. Právě naopak – v zájmu specialistů SEO je, aby uživatelé pokládali jen základní obecné dotazy, vůči kterým lze webové stránky snadno optimalizovat.

Podrobné analýze vyhledávacích strojů primárně z pohledu uživatele, rešeršéra, se tedy věnuje relativně menší (ve srovnání se SEO) komunita informačních specialistů. Ti se zaměřují přímo na konkrétní rozbory fungování pokročilých funkcí a operátorů, zkoumají plné možnosti využití nově uváděných služeb, porovnávají schopnosti jednotlivých vyhledávačů.

Své poznatky pak zachycují na svých osobních stránkách (již zmíněný Notess [3], Bradley [4]), v podobě blogů (Lenssen [5]) nebo je publikují formou elektronického seriálu (Blakemanová [6]). V některých případech spravují tito specialisté celé zpravodajské portály – nejznámějším je pravděpodobně Search Engine Watch [7], vytvořený Danny Sullivanem, od konce roku 2006 šéfredaktorem webu Search Engine Land [8]. Mezi dalšími lze zmínit portály Pandia [9] či ResourceShelf [10]. Pravděpodobně není jen shodou okolností, že mnoho odborníků z této oblasti působí či působilo v knihovnickém prostředí (Notess, Bradley, Price).

Hlavními informačními zdroji této práce jsou výsledky vlastních provedených analýz, ověřované také s pomocí uvedených zdrojů. V některých případech bylo naopak využito výsledků analýz z těchto zdrojů, které byly ověřovány vlastním výzkumem.

1.2 Vymezení pojmů

Vyhledávač, vyhledávací stroj (angl. *search engine*) – aplikace, schopná vyhledávat v textech velkého počtu dokumentů pomocí předběžné indexace jejich obsahu. V rámci této práce je termín „vyhledávač“ používán v širším smyslu i pro webová sídla, která užívání této aplikace pomocí dialogového rozhraní umožňují.

Katalog (angl. *directory*) – v prostředí českého internetu ustálený termín pro rozsáhlou databázi odkazů na webová sídla. Tato databáze je strukturně členěna do kategorií, zpravidla podle obsahového, případně geografického zaměření jednotlivých webových sídel. Pro původní anglický termín *directory* je v této práci používán výhradně termín katalog. Další možné české překlady (adresář, rejstřík, rozcestník, předmětově orientovaný zdroj) nejsou použity, neboť nevystihují zcela srozumitelně podstatu pojmu.

Robot (též *spider*, *crawler*, *bot*) – program, shromažďující pro vyhledávač data z obsahu webových dokumentů. Roboti periodicky procházejí prostřednictvím hypertextových odkazů velké množství webových sídel a zaznamenávají jejich obsah (a změny v obsahu) pro další zpracování.

Dotaz – textový řetězec, vyjadřující formou určité syntaxe informační požadavek uživatele. Dotaz je pokládán vyhledávacímu stroji prostřednictvím dialogového rozhraní. Dotaz se skládá z výrazů a operátorů.

Výraz (též *term*) – základní stavební prvek dotazu. Jednoslovná i víceslovná jazyková konstrukce, reprezentující v dotazu předmět vyhledávání.

Operátor – definuje vztahy mezi výrazy v dotazu, popřípadě blíže upřesňuje vztah výrazu k hledanému dokumentu.

Fráze – stanovená posloupnost více slov. Pokud je výraz zadán jako fráze, budou slova v něm vyhledávána jen v této zadané posloupnosti.

SEO – zkratka pro *Search Engine Optimalization*, česky „optimalizace (stránek) pro vyhledávací stroje“. Soubor činností, prováděných za účelem zviditelnění optimalizované stránky a zvýšení její návštěvnosti.

Tag – speciální textový řetězec, základní prvek, definující strukturu webových dokumentů založených na jazyce SGML. Kromě strukturních informací (elementů) může být nositelem také sémantické informace (metatag).

Nesting (též *hnízdování*) **–** způsob členění složitějších dotazů na dílčí booleovské operace pomocí operátoru závorek.

Stemming – automatické převádění tvarů anglických slov na jejich kmeny odstraňováním koncovek. Stemming v prostředí vyhledávacích strojů slouží jako nástroj pro vyhledání různých tvarů téhož slova. V češtině se z důvodu větší ohebnosti slov k témuž učelu používá proces lematizace.

Google hack – způsob zadávání parametrů dotazu, při němž je místo dialogového rozhraní použito přímé editace URL dotazu.

Zpětný odkaz, příchozí odkaz (angl. *backlink, inbound link*) **–** zpětný odkaz určitého dokumentu je hypertextový odkaz, který na něj odkazuje z jiného webového dokumentu.

Pagerank – Googlem patentovaný způsob hodnocení webových stránek, založený na analýze struktury hypertextových odkazů mezi webovými stránkami.

SERP – zkratka pro Search Engine Result Page, česky „stránka s výsledky vyhledávání“. V oboru SEO představuje koeficient SERP vyjádření úspěšnosti optimalizace. Pokud je hodnota SERP dané webové stránky rovna jedné, je odkaz na tuto stránku přítomen na první straně výpisu výsledků.

1.3 Stručný přehled historického vývoje vyhledávačů

1.3.1 Archie, předchůdce vyhledávacích nástrojů

První nástroj, sloužící pro orientaci v množině dokumentů, dostupných prostřednictvím sítě internet vznikl v roce 1990. Jmenoval se Archie a vytvořil jej Alan Emtage, student montrealské McGill University. V této době neexistoval ještě žádný World Wide Web v dnešní podobě, tedy ani žádné internetové vyhledávače či katalogy. Dokumenty byly dostupné na FTP serverech, ze kterých je uživatelé získávali pomocí FTP klientů. O nově zveřejněných dokumentech se uživatel mohl dovědět především sledováním příspěvků v emailových diskusích.

Program Archie je označován za prvního předchůdcem dnešních vyhledávacích strojů z toho důvodu, že jako první začal fungovat zhruba na podobném principu. Podle určitého seznamu procházel volně přístupné FTP servery a shromažďoval názvy souborů na nich umístěných. V databázi těchto názvů pak mohl uživatel vyhledávat pomocí regulárních výrazů.

Vedle FTP serverů se vyvinuly také Gopher servery, které již obsahovaly čistě jen textové dokumenty. Pro získávání informací o dokumentech na Gopher serverech vznikly v roce 1993 programy Veronica a Jughead, fungující na podobném principu jako Archie.

1.3.2 Počátky World Wide Webu

S nástupem protokolu HTTP začal počet dokumentů, přístupných na vznikajícím World Wide Webu rychle narůstat. Vznikly tak programy, které pro získávání dat začaly využívat specifika HTTP – hypertextových odkazů. Jako první skutečný robot, procházející tímto způsobem prostor WWW je označován World Wide Web Wanderer. Původně mělo procházení sloužit pouze k získávání informací o celkovém počtu webových stránek na síti, posléze však začal robot zaznamenávat také jejich adresu URL. Vznikla tak první databáze webových odkazů zvaná Wandex.

Jako první robot svého druhu měl však Wanderer také své nedostatky. Procházení dokumentů pomocí odkazů nebylo zpočátku nijak omezováno a Wanderer tak v četných případech načítal jednu stránku i několiksetkrát za den. To vedlo ke značnému zpomalování provozu celé sítě. Bylo zřejmé, že kvalita algoritmu,

určujícího cestu robota sítí WWW bude rozhodujícím kritériem pro kvalitu všech dalších budoucích vyhledávačů.

Radikálním způsobem se s tímto problémem vypořádal jiný vyhledávač. Byl jím ALIWEB (Archie-Like Indexing of the Web), který pro shromažďování dat nepoužíval žádnou formu automatického robota, ale naopak od správců webových sídel přijímal ručně vytvářené soubory, ve kterých byla popsána struktura a charakteristika jimi spravovaného webu.

To je velmi odlišný pohled na způsob získávaní dat, specifický svou náročností na lidský faktor. V tomto fundamentálním rozdílu mezi Wandererem a ALIWEBem lze spatřovat prapůvodní základy dvou hlavních budoucích směrů, filozofií, jak uživateli umožnit orientaci ve velkém množství webových dokumentů. Tyto směry reprezentují na jedné straně automatické internetové vyhledávače (search engines) a na straně druhé ručně spravované katalogy (directories).

1.3.3 Rozvoj sítě WWW

Postupný vývoj vyhledávačů založených na práci robota přinesl rychlý růst počtu indexovaných dat v repozitářích těchto vyhledávacích systémů. S tím, jak se rozrůstal i World Wide Web, vyvstaly nové otázky:

- Pro orientaci v narůstající množině webových dokumentů již nestačilo jen vyhledávání podle názvu či URL; bylo nutné začít indexovat alespoň částečně také jejich obsah. To byl problém zejména z hlediska možností tehdejších technických prostředků – diskových kapacit repozitářů, výpočetní doby dotazů, rychlosti přenosových cest.

- Bylo také nutné začít řešit otázku, jakým způsobem řadit výsledky vyhledávání, které nyní již běžně vracelo odkazy na řádově stovky dokumentů. Dosud nebyly výsledky řazeny buď vůbec, nebo jen abecedně. Otázka řazení výsledků podle určité formy relevance k dotazu zůstává problémem dodnes.

Koncem roku 1993 se objevily první vyhledávače, které se snažily uvedené problémy řešit. Robot vyhledávače JumpStation zaznamenával kromě adresy již také obsah záhlaví (metatag `<header>`) webových dokumentů. Výsledky dotazů však

stále nijak neřadil. Prvním vyhledávačem, který určitým způsobem řadil výsledky vzhledem k danému dotazu, byl Repository-Based Software Engineering (RBSE).

1.3.4 První katalogy

Počátkem roku 1994 vznikly také první systémy, nabízející vyhledání dokumentů nikoli pomocí pokládání dotazů, ale postupným procházením skupin hypertextových odkazů, členěných do tematických kategorií a podkategorií. Tímto způsobem vznikly první katalogy, Galaxy a především Yahoo!.

Tento nejznámější katalog původně žádným katalogem nebyl – vznikl jen jako seznam oblíbených webových stránek vytvořený dvěma postgraduálními studenty Stanfordské univerzity v Kalifornii. Když David Filo a Jerry Yang zjistili, že jejich seznam je navštěvován denně tisícovkami uživatelů, začali jej rozšiřovat a dále strukturovat. Výhodou takto lidskými silami spravovaného systému bylo množství dodatečných informací o indexovaných dokumentech.

1.3.5 První plnotextový vyhledávač

Také další přelomový nástroj vznikal původně jako studentský projekt – vyhledávač WebCrawler. Brian Pinkerton, student Washingtonovy univerzity vytvořil vyhledávací stroj, jehož cílem měla být indexace plného textu procházených dokumentů. Vyhledávač si rychle získal značnou popularitu mezi uživateli, avšak jeho provoz byl stále náročnější na hardwarové prostředky celé univerzity. Posléze byl zakoupen komerční firmou America Online a stal se základem pozdějšího vyhledávače Excite.

1.3.6 Altavista

WebCrawler byl jen první z celé řady dalších plnotextových vyhledávačů. V polovině devadesátých let nadále vznikaly nové internetové vyhledávací služby jako Infoseek, Inktomi, Lycos nebo Excite. Jejich vývojem se nyní stále častěji zabývaly komerční subjekty, schopné zabezpečit dostatečně výkonné technické zázemí pro shromažďování velkého objemu dat a pro rychlé zpracovávání uživatelských dotazů.

Jedním z vyhledávačů tohoto typu byla i Altavista, která si držela neoficiální titul nejpopulárnějšího vyhledávače až do nástupu Googlu koncem devadesátých let. Altavista byla výrazná mimo jiné několika novinkami, které se od té doby staly nedílnou součástí takřka každého vyhledávače.

Mezi ně patří:

- schopnost zpracovávat i dotazy položené prostřednictvím přirozeného jazyka
- zavedení booleovských operátorů
- vyhledávání jen v archivech elektronických diskusních skupin či v názvu obrázků
- zavedení operátoru pro zkoumání okolí stránky (`link:`)
- dynamické zobrazování rad pro formulaci dotazů v uživatelském prostředí

1.3.7 Google

Z hlediska popularity i z hlediska obchodního nejúspěšnější internetový vyhledávač všech dob začal vznikat v roce 1996 na Stanfordské univerzitě, stejně jako o dva roky dříve katalog Yahoo!.

Google, podobně jako Yahoo!, byl také původně jen školním projektem dvou studentů, Larryho Page a Sergeje Brina. Oba se zabývali problematikou posuzování objektivní hodnoty velkého počtu webových dokumentů. Řešením byla podle nich podrobná analýza příchozích odkazů, vedoucích na každý dokument. Výsledkem této analýzy byl koeficient nazvaný PageRank, jehož tvorba je popsána ve výzkumné zprávě „The PageRank Citation Ranking: Bringing Order to the Web“ [11].

Page s Brinem na této myšlence založili celý projekt vyhledávacího stroje, který popsali ve své práci „The Anatomy of a Large-Scale Hypertextual Web Search Engine“ [12]. Tak vznikl Google, ačkoli ve svých počátcích se tento pokusný vyhledávač nazýval BackRub, v narážce na analýzu zpětných odkazů (backlinks).

2 Mechanismus fungování vyhledávače Google

V této kapitole budou stručně charakterizovány činnosti a procesy, probíhající v pozadí celého systému vyhledávání.

2.1 Shromažďování dat

Přesná pravidla, jimiž se jeho robot (roboti) řídí při procházení webového prostoru Google pochopitelně opět tají. Není tak zcela jisté, které faktory a do jaké míry ovlivňují trasu robotů v síti WWW. Experimenty provedené v této oblasti nepřinesly jednoznačné závěry [13].

Google používá pro shromažďování dat dva typy robotů – tzv. *deep boty* a *fresh boty*. Deep boti jsou hlavními nástroji Googlu pro získávání dat – procházejí opakovaně celý (v rámci svých možností, viz dále) webový prostor, ukládají datový obsah navštívených webových sídel, případně zaznamenávají informace o změnách v tomto obsahu. Vyhodnocování informací získaných deep boty je hardwarově náročné a obměna hlavního indexu proto probíhá zhruba jen jednou do měsíce.

K průběžným aktualizacím indexu používá proto Google ještě fresh boty. Jejich primárním úkolem není stahování nových dat, ale pouze průběžné zachycování změn v obsahu určité skupiny stránek. Tyto stránky jsou vybírány pravděpodobně podle významnosti (vyjádřené určitou minimální hodnotou PageRanku) a frekvence změn. Úprava hlavního indexu na základě informací od fresh botů probíhá v podstatě neustále, což se projevuje i ve výsledcích vyhledávání, které se tak mohou rychle měnit. V oboru SEO jsou tyto změny nazývány jako *Everflux* („věčný pohyb“).

Ani s pomocí velkého počtu deep botů není možné pokrýt více než jen malý díl skutečného rozsahu celého webu (konkrétní odhady se různí). Jeho značná část tak zůstává nejen pro Google skryta v oblasti souhrnně nazývané jako neviditelný web. Do tohoto pro roboty nedostupného prostoru mimo jiné patří:

1. Dokumenty, k jejichž získání je nutné se manuálně přihlásit a vstoupit do jinak nepřístupné části webového sídla. Typicky se jedná o informace poskytované za úplatu, které tvoří největší (a nejhodnotnější) část neviditelného webu.

2. Dynamicky vytvářené dokumenty, které fyzicky v podobě stálých souborů neexistují. Často jsou vytvářeny jako výsledek interakce uživatele s rozhraním webového sídla. Klasickým případem jsou výstupy z nejrůznějších databázových služeb – například výpisy z obchodního rejstříku, jízdní řády, programy kin apod.

3. Dynamicky vytvářený obsah celého webového sídla. Podobné jako v předchozím bodě. Zde se však nejedná o jeden individuální výsledek uživatelského požadavku, ale celý vlastní obsah webu je generován z databáze. Takto je tomu často v případě článků na zpravodajských webech. Jelikož Google dynamicky vytvářené dokumenty buď vůbec neindexuje, nebo jim přiřazuje menší PageRank, správci těchto sídel maskují stránky navenek do podoby klasických souborů pomocí techniky *mod_rewrite* (viz kapitola 7.1.2).

4. Dokumenty běžně přístupné na webu, jejichž zaznamenání robotem vyhledávačů však autor nebo správce webu technickými prostředky zakázal (prostřednictvím metatagu či konfiguračního souboru robots.txt).

5. Teoreticky robot nenajde takové dokumenty, na něž nevede z žádného jiného dokumentu alespoň jediný hypertextový odkaz. Takové dokumenty je sice možné zaregistrovat přímo u provozovatelů vyhledávacích služeb, jejich PageRank je však nadále velmi nízký.

6. Dokumenty, které robot nenavštívil kvůli technickým potížím (výpadek serveru, nevhodná flashová či javascriptová navigace v rámci sídla, neznámý formát souboru).

7. Dokumenty, nacházející se příliš hluboko ve struktuře webového sídla. Roboti procházejí každý web jen do určité hloubky, v závislosti na PageRanku stránek.

8. Informace, které se nacházejí příliš daleko v těle dokumentu. Podle dlouhodobě tradované teorie Google indexuje pouze prvních 101 kB zdrojového kódu každého dokumentu. Bondar [14] však na základě provedeného experimentu uvádí, že Google zaznamenává ve skutečnosti pětkrát větší objem dat.

Vedle těchto případů lze uvést ještě dvě situace, kdy repozitář Googlu sice informaci obsahuje, ale z důvodu malé přesnosti vyhledávání, typické pro Google (viz kapitola 6.3), je informace pro uživatele v podstatě také „neviditelná“:

- Relevantní dokument se nachází ve výpisu výsledků tak daleko, že se k němu uživatel nedostane. Příčinou může být příliš obecně (nebo nepřesně) položený dotaz, případně velmi nízký PageRank daného dokumentu.

- Pokud je nalezeno více dokumentů v rámci jedné domény, Google zobrazí ve výsledcích nejvíce dva, které jeho algoritmus vyhodnotí jako nejrelevantnější. Zobrazení všech nalezených dokumentů (pro všechny domény ve výpisu) je možné pouze pomocí odkazu, který se nachází až na úplném konci výpisu výsledků. Jinou možností jak vypnout toto omezení je Google hack (parametr `filter=0`).

2.2 Zpracování informací, výpočet PageRanku

Shromážděná data jsou zpracována do podoby invertovaného souboru, respektive je pomocí nich aktualizován stávající indexový soubor.

Pro každý dokument je také v systému vypočítávána hodnota jeho koeficientu PageRank. Tento komplexní výpočet se stal základním kamenem úspěchu Googlu jako internetového vyhledávače. Číselná hodnota PageRanku je určitým vyjádřením důležitosti, relativní informační hodnoty dokumentu na základě analýzy jeho citovanosti dokumenty ostatními.

Unikátnost PageRanku však nespočívá v hodnocení prostého počtu příchozích odkazů, tímto způsobem byla oceňována významnost webových stránek již v minulosti. Podstatou celého výpočtu je myšlenka, že nezáleží primárně na množství odkazujících (citujících) stránek, ale na jejich „kvalitě“ – ta je dána opět jejich citovaností dalšími stránkami, a tak stále dál.

Nazíráno z druhého směru, PageRank každé webové stránky je důležitý pro výši PageRanku všech dokumentů, které cituje (pomocí hypertextového odkazu). Stránka tak poskytuje část svého PageRanku, aniž by jej ztrácela, stránkám dalším. Konkrétní velikost této předávané hodnoty závisí mimo jiné na počtu odkazů přítomných na

stránce (čím větší množství, tím se předávaný PageRank „rozmělňuje“) a pravděpodobně také na tematické příbuznosti.

Tento způsob hodnocení, založený na důkladné analýze provázanosti všech indexovaných stránek, vyžaduje vytvoření a neustálou údržbu jakési hypertextové mapy sítě WWW. Brin a Page popisují tuto mapu ve vztahu k PageRanku takto:

„The citation (link) graph of the web is an important resource that has largely gone unused in existing web search engines. We have created maps containing as many as 518 million of these hyperlinks, a significant sample of the total. These maps allow rapid calculation of a web page's "PageRank", an objective measure of its citation importance that corresponds well with people's subjective idea of importance. Because of this correspondence, PageRank is an excellent way to prioritize the results of web keyword searches. For most popular subjects, a simple text matching search that is restricted to web page titles performs admirably when PageRank prioritizes the results.“ [12]

2.3 Vyhodnocování uživatelských dotazů

Výběr dokumentů pro výsledek konkrétního dotazu a také způsob řazení těchto výsledků je ovlivňován velkým počtem utajovaných faktorů (řádově mnoho desítek). Identifikace těchto faktorů a zkoumání míry jejich vlivu je hlavní náplní oboru SEO, s cílem dosáhnout co nejvyšší hodnoty SERP optimalizovaných stránek [15].

Zde budou zmíněny především ty faktory, které se bezprostředně týkají obsahu dokumentu (tzv. *on-page faktory*).

- Vybírány jsou dokumenty, v jejichž obsahu se hledané výrazy vyskytují.
- Velmi důležitým faktorem, který zvyšuje SERP dané stránky, je přítomnost hledaných výrazů v klíčových elementech stránky – v titulku stránky (tag `<title>`) či nadpisech různé úrovně (`<h1>`,`<h2>`…). Výraz je posuzován jako relativně důležitý také pokud je na stránce zvýrazněn tučným písmem.
- Vyhledány mohou být i stránky, v jejichž těle se výraz vůbec nenachází – Google jej nalezl v těle hypertextových odkazů na tuto stránku odkazujících (viz kapitola 7.1.6).

- Relevance dokumentů vzhledem k dotazu je posuzována také podle četnosti, se kterou obsahují hledaný výraz. Dokumenty s častěji se opakujícím výrazem chápe Google jako důležitější. Pokud však tato četnost překročí určitou mez, Google tuto stránku vyhodnotí jako uměle „přeoptimalizovanou“ a odsune ji ve výsledcích do pozadí. Smička uvádí [16] jako optimální hustotu 2–7 %.

- SERP ovlivňuje do jisté míry i vzdálenost výrazu od začátku těla stránky (resp. určitého tagu). Čím dále v těle stránky, tím je výraz chápán jako méně důležitý.

- Při zadání více výrazů do dotazu hraje roli i jejich vzájemná blízkost v těle prohledávaných dokumentů. Jelikož však Google nedisponuje pokročilým operátorem pro specifikaci vzdálenosti, nelze tuto blízkost stanovit jako rozhodující pro výběr dokumentů.

- Dokumenty obsahující výrazy ve tvaru, v jakém byly zadány do dotazu jsou považovány za relevantnější než dokumenty obsahující pouze jejich odvozeniny.

Nejdůležitějším *off-page faktorem*, tedy takovým, který bezprostředně nesouvisí s obsahem stránky je samozřejmě již zmíněný PageRank.

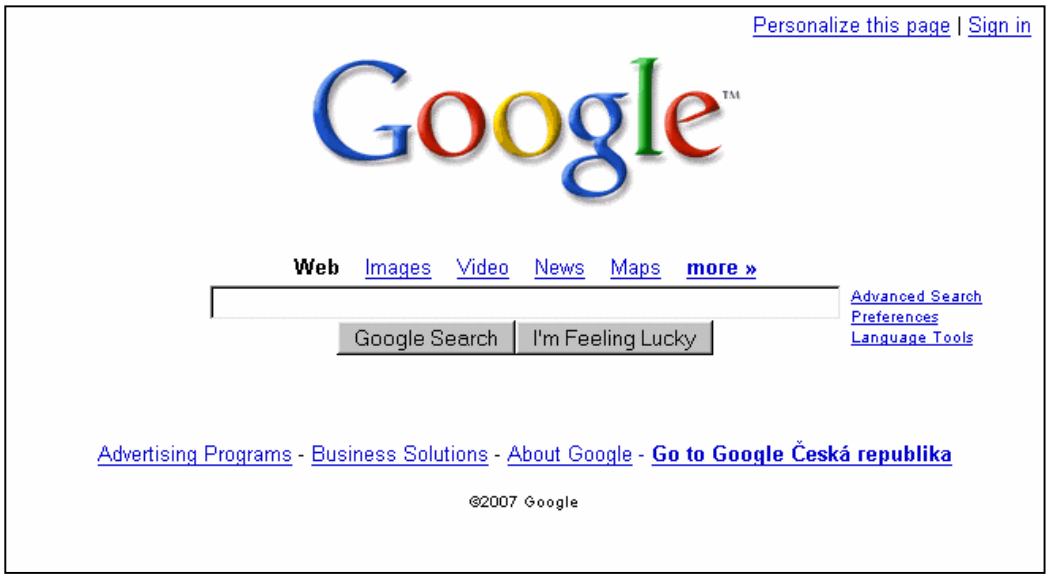
Na závěr této kapitoly je třeba podotknout, že vyhodnocení totožného dotazu může vracet v krátkém časovém rozmezí odlišné výsledky – nejen v důsledku měsíční aktualizace centrálního indexu nebo změny fungování hodnotících algoritmů.

Uživatelské dotazy nejsou totiž zodpovídány vždy fyzicky stejným strojem – Google pro zpracování dotazů používá najednou služeb zhruba osmi mohutných datových center, která má rozmístěna v různých částech světa. Aktualizace či změny v celém systému probíhají postupně. Dva dotazy položené v krátkém časovém úseku tak mohou být zpracovány dvěma různými centry, jejichž funkčnost se může do jisté míry lišit.

3 Analýza uživatelského rozhraní

Nespornou součástí úspěchu Googlu je jednoduchost a přehlednost jeho grafického rozhraní. Strohý styl uživatelského prostředí, zejména titulní stránky, její minimalistický design a prostota se stala nedílnou součástí značky Google. Je také vyjádřením podstaty filozofie Googlu – nezatěžovat uživatele záležitostmi, kvůli kterým nepřišel. Rozhraní má především nabízet co nejpřímější cestu k informacím, které tu uživatel hledá – proto stránku navštívil. Vše ostatní zůstává skryto. Poté, co zadá do jediného přítomného pole svůj dotaz, je prohledáno několik různých databází (hlavní webové vyhledávání, Google News, Google Book Search aj.). Všechny případné výsledky jsou uspořádány do jediného přehledného výpisu. Nejrůznější možnosti a nastavení jsou uživateli zobrazeny pouze tehdy, pokud si je sám vyžádá nebo mají bezprostřední vztah k výsledkům hledání.

3.1 Domovská stránka



Obr. 1: Titulní strana Google.com

Charakteristický prostý design titulní stránky, ačkoliv odpovídá základní myšlence Googlu, vznikl nicméně náhodou. Okolnosti jeho vzniku a vývoje, občas až úsměvné, jsou nedílnou součástí historie firmy a Google je často zmiňuje ve svých prezentacích pro veřejnost. Stejně tak tomu bylo i loňského července, kdy se na pražské ČVUT konala přednáška Jennifer Gove, analytičky z oddělení Googlu pro výzkum uživatelských zkušeností. Prezentace nazvaná „User Experience at Google“

zde proběhla pod záštitou Czech SIGCHI – české pobočky organizace Special Interest Group on Computer-Human Interaction.

Vytvořením prapůvodní domovské stránky byl pověřen Sergej Brin. Oba zakladatelé Googlu byli především programátoři a Brin se webdesignu nijak zvlášť nevěnoval, neovládal příliš ani kódování v HTML. Šlo mu o to, co nejdříve zprovoznit základní funkci Googlu, tedy vyhledávání. V ranných verzích tak nebylo na stránce ani žádné tlačítko, dotaz se potvrzoval stiskem tlačítka Enter. Údajně i samotný název, varianta matematického termínu *googol*, vznikl špatným přepisem.

Obr. 2: Verze titulní strany Googlu z 11. listopadu 1998. Zdroj: Archive.org

Postupem času, s tím, jak systém pronikal mezi širší veřejnost, vzrůstala potřeba zjišťovat uživatelské návyky a názory – co návštěvníci stránky oceňují, co jim vadí či naopak chybí. V lednu roku 2000 se uskutečnil první takový výzkum, který přinesl několik zajímavých poznatků. Jeden z respondentů například seděl bez hnutí před stránkou celé minuty, neboť čekal, kdy se nahraje zbytek stránky – stránka byla prázdná až příliš. To byl hlavní a prvotní důvod toho, že se na spodku titulní stránky začala zobrazovat značka pro copyright, která měla signalizovat konec stránky. Další podobné podněty měly za důsledek změnu loga nebo úpravu jazyka použitého na stránkách.

Vzhled titulní strany se během let měnil tak, jak informací na ní přibývalo i ubývalo [17], zmizela například i dlouholetá informace o počtu indexovaných stránek. Většina změn ale byla drobnějšího charakteru. Základní linie zůstala zachována – mít

na titulní stránce co nejméně zbytečných prvků. A tak ačkoliv portfolio Googlu zahrnuje řádově desítky dalších služeb, svůj vlastní odkaz na titulní straně mají jen čtyři (Images, Video, News a Maps) – Google je vybírá (zřejmě) podle návštěvnosti.

Ukrytí výkonného vyhledávacího stroje za minimum ovládacích prvků mělo dva důsledky. Za prvé si velmi prostý design získal řadu začátečníků. Noví uživatelé si lépe zvykali, prosté prostředí v nich posilovalo sebedůvěru, že vše na obrazovce mají pod kontrolou a vědí, co k čemu slouží. Druhým důsledkem byl výrazně odlišný vzhled jako konkurenční výhoda. Pro ostatní již zavedené vyhledávací služby bylo často obtížné napodobit uživatelsky přívětivou strohost Googlu, neboť tím byly nuceny k razantnímu osekávání svého členitého rozhraní, zatímco Google začínal „z ničeho".

3.2 I'm Feeling Lucky

Studie uživatelského chování se staly běžnou a pravidelnou součástí dalšího vývoje. Google upřednostňuje testování uživatelů v reálném provozu, často tedy nějakou službu spustí v beta verzi (která může běžet i roky) a upravuje ji podle ohlasů.

Jedním ze zajímavých prvků rozhraní je druhé tlačítko přítomné na titulní straně – „I'm Feeling Lucky" (ve významu „Myslím, že budu mít štěstí", „Risknu to"). Po jeho stisknutí se provede normální vyhledání výsledků, ale uživatel je ihned automaticky přesměrován na stránku, která je ve výsledcích zařazena jako první. Je tedy zcela pominuta možnost hodnocení relevance výsledků uživatelem, tedy něco, co si informační rešeršér nemůže dovolit.

Navíc z textu tlačítka „I'm Feeling Lucky" (v české verzi o něco výstižnější „Zkusím štěstí") řada uživatelů nepochopí jeho funkci. Tlačítko má tak v podstatě téměř nulovou užitnou hodnotu, navíc stejnou funkci automatického přechodu na první výsledek mají v sobě již zakomponovány některé prohlížeče. V Mozille Firefox tak stačí zadat dotaz přímo do adresového řádku a potvrdit – v novějších verzích prohlížeče je zobrazena stránka s výsledky z Googlu, ve starších verzích je přímo načítána první stránka.[1]

[1] V Internet Exploreru se takto dotaz vyhledá v Microsoft Live Search, Opera v české verzi se pokusí přejít na adresu ve tvaru http://www.dotaz.cz, pokud nenajde, pak na http://www.dotaz.com

To, co je na tomto tlačítku opravdu pozoruhodné, je fakt, že Google z výzkumů uživatelského chování ví, že většina uživatelů buď nechápe jeho význam, nebo chápe, ale stejně ho nepoužívá. Přesto jej na titulní straně ponechává. Uživatelé totiž v průzkumech opakovaně označují tlačítko jako prvek, který by na stránce postrádali, jeho přítomnost dodává v jejich očích Googlu trochu humorného rázu a s ním i lidskou tvář.

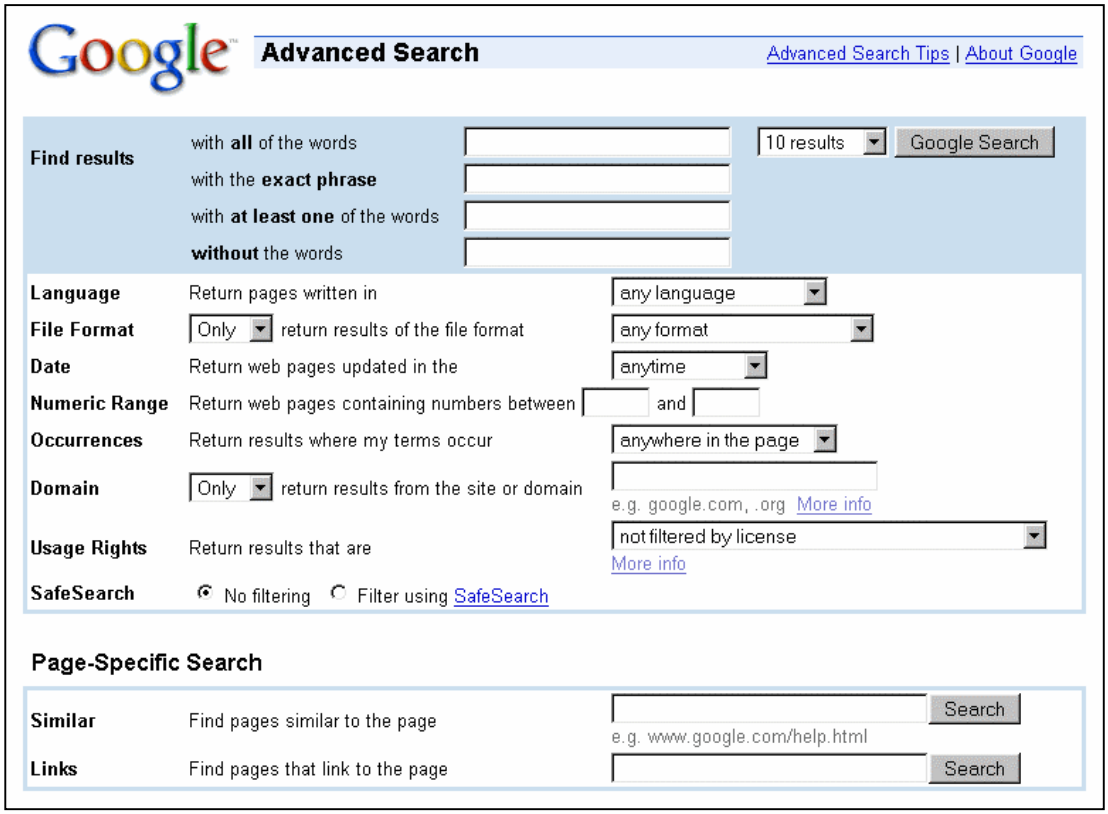
Google si tohoto „polidšťujícího“ faktoru velmi cení a podporuje jej. Mimo jiné i tím, že o různých svátcích, výročích či jiných příležitostech doplňuje své logo různými grafickými doplňky – například jen během olympiády v Turíně obměnil takto své logo dvanáctkrát.

Obr. 3: Logo Googlu u příležitosti dne sv. Patrika, výročí narození Vincenta van Gogha a olympiády v Turíně. Zdroj: Google.com

O tom, že Google bere humor jako svou přirozenou součást svědčí i následující skutečnost – při výběru národní verze uživatelského rozhraní může uživatel vybírat i z takových, jejichž jazykem žádný skutečný národ pravděpodobně nemluví. Jsou jimi: klingonské rozhraní (podle mimozemské civilizace z televizního seriálu Star Trek), elmer-fudské a „Bork, bork, bork!“ (inspirované postavami z televizních pořadů pro děti), hackerské (užívající záměn písmen za čísla) a pig-latinské (podle jazyka založeného na jazykové hříčce).

Všechny zmíněné prvky, které zdánlivě ubírají vyhledávači na serióznosti, však nenásilně posilují jeho uživatelsky přívětivý image. Tedy něco, co ve sféře internetových vyhledávacích služeb byla a je výrazná konkurenční výhoda.

3.3 Advanced Search



Obr. 4: Obrazovka pokročilého vyhledávání

Formulář pro pokročilé vyhledávání nabízí většina velkých vyhledávacích služeb a Google není výjimkou. Ve formuláři jsou k dispozici vkládací textová pole pro každý zde vysvětlený operátor. Pomocí těchto polí je tedy možné vytvořit kombinovaný dotaz, aniž by uživatel znal syntaxi operátorů. To může být velmi vhodný způsob jak zadávat složitější dotazy zejména pro začínající či občasné uživatele. Nejen z hlediska informačního specialisty je však dobré si být vědom určitých omezení, které toto rozhraní má:

- Samotná struktura formuláře (jedno pole na jeden operátor) neumožňuje některé operátory použít v dotazu vícekrát, konkrétně formát souboru (filetype), číselný interval (numeric range) a omezení domény (site).
- Většímu kombinování brání i fakt, že v poli Occurences, určujícím pozici hledaných výrazů v rámci stránky, lze zvolit pouze jednu možnost (tedy vlastně jen jeden z operátorů `allintitle:`, `allintext:`, `allinurl:`, `allinanchor:`).

- Tento vybraný operátor se vztahuje na všechny výrazy v dotazu bez rozdílu (předpona „all“) – nelze omezit jeho platnost jen na některé z nich.

- Výběr jazyka hledaných stránek (Language) a nastavení počtu vypisovaných výsledků na stránku se váže jen ke konkrétnímu jednomu dotazu, nezůstává trvalým nastavením (narozdíl od nastavení ve složce Preferences).

Naopak některé vlastnosti dotazu lze nastavit pouze v tomto formuláři:

- Date – omezení výsledků na dokumenty, jejichž vznik (či změnu) Google zaznamenal během poslední doby.

- SafeSearch – filtrace stránek s výrazně sexuální tematikou, odkazuje na stejné nastavení v sekci Preferences.

- Usage Rights – specifikace autorských práv vázaných k dokumentu (prostřednictvím licence Creative Commons)

Tyto funkce nemají vlastní operátor a jediným dalším způsobem jejich použití je tzv. *Google hack*, neboli přímá úprava parametrů v URL adrese dotazu. V případě omezení datem pak Google hack skýtá i větší možnosti (přesnější nastavení).

Existují i jiná rozhraní, vybudovaná nad vyhledáváním Googlu třetími stranami. Jsou jimi například Google Ultimate Interface[1], orientovaný na pokročilé uživatele a Soople[2], integrující do jednoho formuláře hledání v různých databázích Googlu (Books, Movies, Google Scholar aj.). Největší volnost však pro informačního rešeršéra stále představuje ruční tvorba dotazu pomocí operátorů.

3.4 Preferences

Na rozdíl od předchozí sekce, obrazovka uživatelských preferencí umožňuje dlouhodobé nastavení voleb (ukládá se do cookies), které ovlivňují všechny položené dotazy. Důležité jsou z hlediska výsledků zejména tyto:

[1] http://www.faganfinder.com/google.html
[2] http://www.soople.com

- Interface Language – jazyková verze rozhraní. Ve skutečnosti ovlivňuje mnohem více, než jen jazyk, kterým Google komunikuje s uživatelem. Jeho nastavení má zásadní vliv na samotné výsledky dotazu (viz kapitola 4).

- Search Language – jazyk hledaných dokumentů. Aktivuje přepínací dvoustavové tlačítko (radio buton), kterým se poté dá jakýkoliv dotaz směrovat buď do oblasti všech dostupných dokumentů, nebo jen dokumentů ve vybraném jazyce.

- SafeSearch Filtering – odstraňuje z výsledků dokumenty s otevřenou sexuální tematikou

3.5 Language Tools

Sekce jazykových nástrojů je dostupná pouze z titulní strany, po položení dotazu se v rozhraní Googlu zobrazují odkazy již jen na Advanced Search a Preferences, což její využívání poněkud stěžuje. Sekce obsahuje následující, poněkud nestejnorodou směs nástrojů.

- Znovu je zde přítomno nastavení jazyka hledaných dokumentů, nejedná se však o trvalé nastavení jako v případě sekce Preferences.

- Naopak pouze zde lze nastavit omezení státu, ve kterém se nalezené stránky mají nacházet. Toto poněkud vágní vymezení je důležité zejména pro anglofonní uživatele, kterým pomáhá vzájemně odlišit stránky například ze Spojených států, Austrálie, Spojeného království apod. Google se snaží odhadnout příslušnost dané stránky k určité geografické poloze zejména podle jejího jazyka, domény 1. řádu a IP adresy serveru, na kterém je stránka hostována [18]. Alternativními způsoby výběru cílové země je použití Google hacku a také hledání přes příslušné regionální verze Googlu (např. www.google.co.uk).

- Další užitečnou pomůckou je interní překladač Googlu, který dokáže překládat vložený text vzájemně mezi hlavními světovými jazyky (v současné době je ve zkušebním provozu též překládání z asijských jazyků). Ještě silnějším nástrojem je možnost překládat celé stránky – po vložení adresy Google zobrazí celou stránku v jejím původním grafickém vzhledu, jen s texty přeloženými např. do

angličtiny. Zároveň zachová funkční všechny odkazy, takže je možné z dané stránky procházet před odkazy dále, stále v anglickém jazyce.

- Okno „Use the Google Interface in Your Language“ nabízí paletu různých verzí rozhraní Googlu v jednotlivých jazycích. Opět se ale jedná jen o prosté odkazy, nejedná se o nastavení trvalé, které je specifikováno v sekci Preferences.

- Ke zmatení uživatele přispívá také poslední oddíl „Visit Google's Site in Your Local Domain“, který obsahuje odkazy na lokální verze Googlu v různých národních doménách. Na rozdíl od okna předchozího, kde byly nabízeny různé jazykové verze jediné stránky www.google.com jsou toto však fyzicky zcela separátní stránky. Google je vytvořil přímo v jazyce a doménovém prostoru daných států, mají stejný vzhled a používají stejný vyhledávací stroj, nicméně se nejedná o identické stránky.

Obě verze se odlišují funkčně i graficky jak od verze anglické, tak od sebe navzájem. Nejsnáze je lze odlišit buď dle jejich adresy URL, nebo vizuálně podle podoby loga. Zatímco česká verze hlavního serveru Google.com má v logu text „Google česky“, lokální Google.cz se jmenuje „Google Česká republika“.

Problematice funkčních rozdílů mezi těmito opticky totožnými lokálními verzemi se věnuje kapitola 4.

3.6 Česká verze uživatelského rozhraní

V českém programovém prostředí je nový návštěvník serveru www.google.com vždy přesměrován na jeho lokální, českojazyčnou verzi www.google.cz. Tak se děje do té doby, než zde klikne na odkaz „Google.com in English“, čímž se v prohlížeči vytvoří cookie z této centrální domény a přesměrování ustane.

Ačkoliv běžný uživatel rozhraní v českém jazyce přivítá a pravděpodobně si ani neuvědomí, že se nachází na jiném serveru (v případě Google.cz), v českém Googlu chybějí některé prvky z anglického rozhraní:

- Uživatel nemá možnost využít velké části dalších vyhledávacích služeb Googlu – v anglickém rozhraní dostupných pomocí odkazu „More“ nad vyhledávacím polem. Taktéž nabídka služeb umístěná nad tímto polem je odlišná.

- V případě serveru Google.cz chybí také možnost přihlásit se k uživatelskému účtu (Google account), česká verze Google.com přihlášení umožňuje.

- V okně pokročilého vyhledávání chybí možnost zadat operátor číselného intervalu a specifikaci autorských práv (Usage Rights).

- V české verzi nelze aplikovat filtr SafeSearch, neboť ten byl vyvíjen jen pro anglický jazyk.

- V jazykových nástrojích chybí okno pro překlad, které by mohl využít i český uživatel, byť třeba jen k překladu z cizích jazyků do angličtiny.

Rozhraní lokálních „poboček“ Googlu jsou sice postupně aktualizována a doplňována, děje se tak ale z velké části překvapivě jen na bázi dobrovolných překladů uživatelů. Tím se dají vysvětlit nejen chybějící prvky, ale i jiné jazykové nesrovnalosti (například odkaz „Pokročilé vyhledávání“ na titulní straně vede do sekce s názvem „Rozšířené vyhledávání“).

Největším rozdílem českého prostředí však není jeho grafická podoba, ale jeho funkčnost, jak dokumentuje následující kapitola.

4 Analýza odchylek ve vyhledávání v závislosti na jazykových nastaveních

4.1 Problematika

Většina uživatelů se domnívá, že výsledky vyhledávání Googlu ovlivňuje pouze obsah (respektive formulace) zadávaného dotazu. Část z nich si je vědoma, že roli hraje i to, zda budou vyhledávat v prostoru celého webu („Prohledat Web“) nebo omezí hledání na stránky v českém jazyce („Stránky pouze česky“). Co však pravděpodobně tuší jen malá část uživatelské veřejnosti je skutečnost, že Google bere v potaz několik dalších faktorů, pomocí nichž vybírá a pořádá výsledky. Tyto skrytě působící faktory jsou dva: **jazyk rozhraní** a **výběr serveru.**

V důsledku toho tak hraje značnou roli to, zda je dotaz položen v českém, nebo anglickém jazykovém prostředí Googlu. Stejně tak různé výsledky dodá i dotaz položený na serveru Google.com a jiné tentýž dotaz (ve stejném jazykovém prostředí) na Google.cz. Výsledky se v obou případech liší tím, že při volbě českého rozhraní, resp. českého serveru jsou do určité míry upřednostňovány české výsledky před výsledky ostatními. Neplatí to ovšem vždy a vždy stejně; z důvodu značné komplexnosti (a utajení) vyhledávacího algoritmu nelze vyvozovat jednoznačné závěry, vybrány a upřednostněny mohou být různé (i nečeské) výsledky.

Tato problematika není samozřejmě jen záležitostí češtiny, ale všech lokálních verzí Googlu. Dokonce i v jazykově jednolitém anglofonním prostředí mají uživatelé na lokálních anglických serverech Googlu (Google.co.uk, Google.com.au) podobné potíže. Problém se zde pouze přesunul z roviny upřednostnění jazyka do roviny upřednostnění místních (britských, australských…) výsledků. Zde je vše ještě složitější, protože odlišení stránek dle místa, ke kterému se váží, je pro Google značně obtížnější než odlišení jen podle použitého jazyka. Používá k tomu nepřímé indicie jako je národní doména, v jejímž prostoru se stránka nachází a IP adresy serveru, na kterém se fyzicky nalézá [18].

Výrazněji se začaly rozdíly ve výsledcích jinak identických dotazů projevovat zhruba v roce 2005 [19]. Ačkoliv pro běžného „lokálního“ uživatele se může jednat o příjemné zvýšení uživatelského pohodlí, z pohledu práce informačního profesionála

jde o nevítanou překážku. Systém takto totiž vyzdvihuje lokální výsledky i v případech, kdy o to není výslovně žádán (navíc při zadání „lokálně nezatížené“ adresy www.google.com je český návštěvník standardně přesměrován na české rozhraní českého serveru www.google.cz). Proto bude v této kapitole proveden výzkum, jehož cílem bude vyčíslit míru vzájemné odlišnosti (inkonsistence) těchto výsledků v různých situacích.

4.2 Metodika testování

Kombinací výše zmiňovaných dvou faktorů spolu s nastavením jazyka hledaných stránek vzniká (pokud uvažujeme jen češtinu a angličtinu) 12 variant téhož vyhledávače a hypoteticky tedy i 12 sad výsledků. Mnoho z nich jsou však varianty ne zcela praktické a v praxi užívané, proto bylo pro potřeby testu vybráno jen šest:

#	server	rozhraní	jazykové omezení	parametry URL
1	Google.com	anglické	jen anglické stránky	hl=en&lr=lang_en
2	Google.com	anglické		hl=en
3	Google.com	české		hl=cs
4	Google.cz	české		hl=cs
5	Google.com	české	jen české stránky	hl=cs&lr=lang_cs
6	Google.cz	české	jen české stránky	hl=cs&lr=lang_cs

Tab. 1: Testované varianty

Poslední sloupec tabulky uvádí parametry URL, pomocí nichž lze nastavení jazyka rozhraní a jazykového omezení hledaných stránek nejjednoznačněji určit a také nejrychleji měnit.

Je předem zřejmé, že výsledky z variant 1, 5 a 6 se budou od zbytku lišit, jelikož explicitně omezují okruh cílových dokumentů dle jazyka; zahrnuty však nebyly z důvodu dokazování jejich odlišnosti, ale proto, aby dokreslovaly odlišnosti zbylých tří. Jedná se v podstatě o jazykově mezní případy – varianta 1 reprezentuje „nejangličtější“ způsob hledání, varianta 6 ten nejvíce český. Zbylé varianty se pak budou pohybovat v prostoru mezi těmito dvěma extrémy.

Testovací dotazy budou postupně pokládány v každé z šesti verzí. Pomocí Javascriptu vytvořeného pro potřeby výzkumu budou z výsledkové stránky extrahovány URL adresy prvních třiceti výsledků. Tento počet byl zvolen záměrně,

neboť při standardním nastavení Googlu odpovídá prvním třem stránkám výpisu výsledků. To je prostor, v kterém si podle průzkumů vybírá výsledky velká většina uživatelů (88% dle výzkumu společnosti iProspect [20]).

Extrahované adresy budou následně v prostředí Microsoft Excelu porovnávány s adresami získanými z ostatních pěti variant. Pro zachycení skutečných faktických rozdílů mezi testovanými třicítkami výsledků budou pominuty rozdíly v jejich řazení, porovnáván bude pouze jejich obsah. Celkovým výstupem analýzy každého jednoho dotazu bude tedy procentuální tabulka o rozměrech 6x6 políček.

4.3 Faktor diakritiky

Ještě před samotným porovnáváním variant je nutné zmínit další zásadní faktor, který je od sebe odlišuje a který hraje roli z hlediska českého uživatele – rozdílný přístup k diakritice. Za účelem dosažení co nejvyšší úplnosti vyhledávání Google v dotazech standardně ignoruje nejen interpunkci, psaní malých a velkých písmen, ale také diakritiku. To má zajisté své výhody a nevýhody, rozhodující (a ne zcela zjevné) je však to, že tak činí pouze při nastavení anglického rozhraní.

Ať už dotaz obsahuje jakoukoliv diakritiku, v anglickém rozhraní jsou vyhledány nejen výsledky se zadaným tvarem výrazu, ale také výrazy zcela bez diakritiky a také se zcela rozdílnou diakritikou! Oproti tomu české rozhraní tak činí pouze v případě, že byl dotaz bez diakritiky zadán. Pokud se nějaká ve výrazu vyskytuje, je hledán jen tento jediný tvar. Situace je znázorněna v tabulce na příkladu slov `plast` a `plášť`.

<table>
<tr><th>dotaz</th><th></th><th>anglické rozhraní (varianty 1,2)</th><th>české rozhraní (varianty 3–6)</th></tr>
<tr><td>"plášť"</td><td rowspan="8">vyhledá dokumenty obsahující</td><td colspan="2">pouze plášť</td></tr>
<tr><td>"plast"</td><td colspan="2">pouze plast</td></tr>
<tr><td>plast</td><td colspan="2">různé varianty (plast, plást, plášť…)</td></tr>
<tr><td>plášť</td><td>různé varianty</td><td>pouze plášť</td></tr>
<tr><td>plást</td><td>různé varianty</td><td>pouze plást</td></tr>
<tr><td>plast -plast</td><td colspan="2">různé varianty mimo tvar plast</td></tr>
<tr><td>plášť -plášť</td><td>různé varianty mimo tvar plášť</td><td>žádné výsledky</td></tr>
</table>

Tab. 2: Role diakritiky v testovaných variantách

Dotazy "plášť" a "plast" ukazují nejbezpečnější možnost, jak Google přinutit hledat pouze výraz v zadaném tvaru. K tomu jsou použity uvozovky jinak užívané jako operátor pro frázi. Ekvivalentní výsledky by též přinesl dotaz +plášť, resp. +plast.

Zadání dotazu plast bez „ochranných“ uvozovek a bez diakritiky vede k největšímu počtu výsledků, ve všech variantách rozhraní jsou vyhledávány všechny jeho různé tvary.

V případě dotazů plášť, plást dochází k již zmiňovanému rozkolu. Zatímco anglické varianty 1 a 2 budou hledat více různých tvarů a nejvíce v popředí se umístí výsledky s mezinárodním termínem „plast“, české budou diakritiku dodržovat a zobrazí tak na předních pozicích zcela jiné výsledky.

Poslední dva řádky tabulky dokládají, že i zdánlivě nesmyslné dotazy mohou mít při určování diakritiky význam.

4.4 Srovnání variant v modelových dotazech

4.4.1 Dotaz "páteř"

Pro demonstraci budou nejprve srovnávány odlišnosti na příkladě jednoduchého, pouze jednoslovného modelu. Jako vhodný příklad polysémie byl zvolen dotaz páteř, protože nabývá různých významů nejen v češtině (v závislosti na diakritice – páteř, páter, pater), ale i v cizích jazycích (zejména latinské pater, „otec“). Prvním analyzovaným případem je tedy dotaz "páteř", který představuje čistě český výraz, opatřený uvozovkami, takže ve všech šesti testovaných variantách bude vyhledáván pouze tento tvar.

Hodnoty v tabulce vyjadřují procento shodnosti prvních třiceti výsledků mezi příslušnými dvěma variantami. Varianty jsou kromě čísla identifikovány i zkráceným popisem ve tvaru server/jazyk rozhraní/jazykové omezení. Pro lepší orientaci jsou výsledky barevně odlišeny, čím intenzivnější odstín červené, tím větší shoda.

"páteř"							
varianta		1	2	3	4	5	6
1	.com /en /en		7	0	0	0	0
2	.com /en	7		87	87	87	87
3	.com /cs	0	87		100	100	100
4	.cz /cs	0	87	100		100	100
5	.com /cs /cs	0	87	100	100		100
6	.cz /cs /cs	0	87	100	100	100	

Tab. 3: "páteř" – procentuální shoda výsledků

Výsledky je možné pro přehlednost zobrazit též v redukované formě, kdy jsou znázorněny pouze vztahy mezi sousedními variantami, které k sobě mají jazykově nejblíže. Navíc lze vizuálně sloučit i varianty, které jsou totožné nejen co do shody zkoumaných třicítek, ale i dle celkového počtu nalezených dokumentů.

"páteř"							
varianta	1		2		3 4		5 6
výsledků	288	7	391 000	87	390 000	100	382 000

Tab. 4: "páteř" – redukované srovnání sousedních variant

Na tento diakriticky jednoznačný a jednojazyčný dotaz Google odpověděl třemi sadami českých výsledků. Všechny varianty s českým rozhraním dodaly v rámci vzorku totožné výsledky. Varianta 2 má také velmi podobné, ale ne zcela shodné výsledky. Pozoruhodné jsou výsledky z varianty 1, neboť by se zdálo logické, že na anglických stránkách v anglickém jazyce se nebude české slovo páteř vyskytovat vůbec. Google přesto našel 288 vyhovujících dokumentů. Ty z určitého důvodu považuje za anglické, ačkoliv přinejmenším celý zkoumaný vzorek tvořily stránky pouze v českém jazyce.

4.4.2 Dotaz páteř

Tentýž dotaz, ovšem bez uvozovek, nezmění nic na českých variantách (3 až 6), stále budou zobrazovat tytéž výsledky s výrazem páteř. Ovšem první dvě varianty, které se díky anglickému rozhraní diakritikou neřídí, naleznou nyní mnohem více dokumentů. Takových, které obsahují jakýkoli diakritický tvar slova páteř – v celosvětové konkurenci se však na předních pozicích umístí jen stránky s nejběžnějším mezinárodním výrazem pater.

páteř							
varianta		1	2	3	4	5	6
1	.com /en /en		100	0	0	0	0
2	.com /en	100		0	0	0	0
3	.com /cs	0	0		100	100	100
4	.cz /cs	0	0	100		100	100
5	.com /cs /cs	0	0	100	100		100
6	.cz /cs /cs	0	0	100	100	100	

páteř							
varianta	1		2		3	4	5	6
výsledků	1,2 mil.	100	8,36 mil.	0	392 000	100	383 000	

Tab. 5 a 6: `páteř` – procentuální shoda výsledků a redukované srovnání

4.4.3 Dotaz `pater`

pater							
varianta		1	2	3	4	5	6
1	.com /en /en		97	13	13	0	0
2	.com /en	97		17	17	0	0
3	.com /cs	13	17		100	67	67
4	.cz /cs	13	17	100		67	67
5	.com /cs /cs	0	0	67	67		100
6	.cz /cs /cs	0	0	67	67	100	

Tab. 7: `pater` – procentuální shoda výsledků

Především při tomto dotazu jsou odlišnosti patrné nejvíce – různé diakritické tvary slova `pater` budou nyní vyhledány už i českými variantami (3 až 6). Protože se již nejedná o výlučně český výraz, začnou se projevovat rozdíly také mezi nimi navzájem – roli začne hrát výběr jazyka dokumentu.[1] Anglické varianty 1 a 2 budou vykazovat stále stejné výsledky (obsahující zejména latinské pater či Pater jako příjmení osoby). Výsledky verzí 3 a 4 budou obsahovat dokumenty v různých jazycích, v 5 a 6 budou pouze české.

[1] Byly zjištěny dokonce i rozdíly mezi servery Google.com a Google.cz, zde se však jednalo jen o drobné změny v pořadí výsledků.

Pro ilustraci rozdílů je v následující redukované tabulce zachycena i jazyková skladba nalezených dokumentů.

pater								
varianta	**1**		**2**		**3**	**4**		**5**	**6**
výsledků	1,2 mil.	97	8,5 mil.	17	8,07 mil.		67	1,25 mil.	
jazyk	29x anglický 1x maďarský		28x anglický 1x maďarský 1x holandský		21x český 4x holandský 4x anglický 1x německý			30x český	

Tab. 8: `pater` *– redukované srovnání sousedních variant*

4.4.4 Dotaz "pater"

"pater"							
varianta		**1**	**2**	**3**	**4**	**5**	**6**
1	.com /en /en		97	13	13	0	0
2	.com /en	97		17	17	0	0
3	.com /cs	13	17		100	60	60
4	.cz /cs	13	17	100		60	60
5	.com /cs /cs	0	0	60	60		100
6	.cz /cs /cs	0	0	60	60	100	

Tab. 9: `"pater"` *– procentuální shoda výsledků*

Dotaz `"pater"`, tentokrát znovu v uvozovkách, vyhledává pouze tento jeden tvar. Ve variantách 1 a 2 nedojde tedy v podstatě k žádným změnám, u ostatních pak ano – ačkoli hodnoty rozdílů mezi jednotlivými variantami zůstávají takřka stejné. Proto jako důkaz byla redukovaná tabulka doplněna o vertikální srovnání s předchozím dotazem `pater`.

pater	**1**		**2**		**3**	**4**		**5**	**6**
	1,2 mil.	97	8,5 mil.	17	8,07 mil.		67	1,25 mil.	
	100		100		50			30	
"pater"	**1**		**2**		**3**	**4**		**5**	**6**
	1,2 mil.	97	8,41 mil.	17	8,04 mil.		60	946 000	

Tab. 10: pater a "pater" – redukované srovnání sousedních variant a dotazů

4.5 Srovnání variant při běžných dotazech

Předchozí modelové testy měly sloužit především k demonstraci možných odlišností jednotlivých variant, jejich výsledky není možné v žádném případě příliš zobecňovat. Ne všechny výrazy jsou tak mnohoznačné jako zvolené slovo `páteř`. Uživatelé také málokdy zadávají dotazy jednoslovné, navíc takto obecné. Taktéž nasměrování českých dotazů do anglicky psaného webu (a naopak) se zkušenější uživatel vyhne. Ale především: největší úskalí nespočívá ve vyhledávání jazykově vyhraněných dotazů – u nich uživatel může očekávat a také obdrží výsledky zkreslené jazykem dotazu.

Problém představují mezinárodní výrazy, vyhledávané v prostředí celého webu, tedy ve variantách 2 a 3. Typickým příkladem jsou jména osob, názvy institucí, ale také čistě anglické, leč do češtiny nepřekládané pojmy, kterých neustále přibývá (např. GSM banking, public relations, sales management aj.). U nich uživatel neklade důraz na jazyk, spíše očekává, že dostane obecně ty nejrelevantnější možné výsledky. Nemusí vůbec tušit, že jsou zatížené volbou českého rozhraní (varianta 3). Tím spíše, že si jej ani nemusel explicitně zvolit – zadání adresy Google.com českého uživatele automaticky přesměruje na Google.cz s jeho předvoleným českým rozhraním.

Chování několika takových mezinárodních výrazů (konkrétně jmen) bude proto pomocí stále stejné metodiky otestováno, za účelem zjištění rozdílů především mezi variantami 2 a 3.

4.5.1 Mezinárodní termíny

"Alan Turing"						
	1	**2**	**3**	**4**	**5**	**6**
1		93	53	53	0	0
2	93		**60**	60	0	0
3	53	**60**		100	3	3
4	53	60	100		3	3
5	0	0	3	3		100
6	0	0	3	3	100	

"guerilla marketing"						
	1	**2**	**3**	**4**	**5**	**6**
1		90	50	50	0	0
2	90		**60**	60	0	0
3	50	**60**		100	13	13
4	50	60	100		13	13
5	0	0	13	13		100
6	0	0	13	13	100	

Tab. 11 a 12: `"Alan Turing"` *a* `"guerilla marketing"`

4.5.2 Mezinárodní termíny s vazbou na české prostředí

"Vaclav Havel"						
	1	2	3	4	5	6
1		83	27	27	3	3
2	83		**43**	43	10	10
3	27	**43**		100	60	60
4	27	43	100		60	60
5	3	10	60	60		100
6	3	10	60	60	100	

"Franz Kafka"						
	1	2	3	4	5	6
1		87	20	20	0	0
2	87		**33**	33	0	0
3	20	**33**		100	40	40
4	20	33	100		40	40
5	0	0	40	40		100
6	0	0	40	40	100	

Tab. 13 a 14: `"Vaclav Havel"` *a* `"Franz Kafka"`

4.6 Závěr analýzy

Cílem provedených testů bylo především dokumentovat vzájemné odlišnosti mezi výsledky a vysvětlit jejich původ. Konkrétní hodnoty rozdílů se liší dotaz od dotazu a je samozřejmě nemožné usuzovat z několika příkladů na celek. Nicméně lze z těchto a předchozích srovnání vyvodit přinejmenším tři obecné závěry:

1. Bylo prokázáno, že volba jazyka rozhraní (při zachování jinak stejných podmínek) viditelně ovlivňuje výsledky dotazu a to i v případech, kdy se dotaz skládá z jazykově neutrálních výrazů.

2. Naproti tomu výběr mezi serverem Google.com a Google.cz má v podstatě zanedbatelný vliv na výsledek hledání, resp. nemá vliv na faktický obsah zkoumaného vzorku třiceti výsledků. Změna pořadí výsledků nebyla měřena.

3. Z předchozích dvou závěrů tedy vyplývá skutečnost, že na jeden dotaz je možné v rámci testovaných šesti variant obdržet až čtyři různé sady výsledků. K nim lze dospět těmito způsoby vyhledávání:

    - vyhledávání omezené na stránky v anglickém jazyce
    - vyhledávání omezené na stránky v českém jazyce
    - vyhledávání jazykově neomezené, prováděné v anglickém rozhraní
    - vyhledávání jazykově neomezené, prováděné v českém rozhraní

4.7 Další odchylky lokálních jazykových verzí

Další odchylky se již netýkají samotného složení výsledků, ale spíše doplňkových prvků uživatelského rozhraní. České prostředí (varianty 3 až 6) tak nedisponuje některými funkcemi, primárně určenými pro anglicky mluvící uživatele:

- Logicky je v českém rozhraní neaktivní kontrola pravopisu dotazů, která upozorňuje na pravděpodobné překlepy.

- V anglickém rozhraní je u cizojazyčných neanglických výsledků automaticky zobrazena možnost „Translate this page“, umožňující jejich překlad do angličtiny; v české verzi tato užitečná pomůcka chybí.

- V českém rozhraní je věta uvozující výsledky (např. „Výsledky 1 - 30 z asi 69 200 na dotaz ash beech cyprus.“) prostým textem. Naproti tomu v anglické verzi je každé slovo z dotazu hypertextovým odkazem k jeho slovníkové definici na serveru Answers.com.[1]

Značně nesystematicky také funguje v různých variantách doprovodné začleňování výsledků vyhledávání z jiných databází Googlu – např. z Google Books, Google News, Local Search, z číselného vyhledávání (např. patenty) nebo vyhledávání definic. Zde se překvapivě projevily také rozdíly nejen mezi variantami s různým rozhraním, ale i mezi servery Google.cz (varianty 4 a 6) a Google.com (3 a 5). Rozdíly dokumentuje tabulka 15.

služba	příklad	1	2	3	4	5	6
Google Books	Moby Dick	ano	ano	ano		ano	
Google News	Tony Blair	ano	ano				
Local Search	London hospital	ano[2]	ano[2]	ano		ano	
definice	define:“de iure“	ano	ano	ano	ano	ano	ano
číselné vyhledávání	patent 5123123	ano	ano				

Tab. 15: Srovnání variant z pohledu funkčnosti sekundárních vyhledávání

[1] Nejedná se o totožné definice, jaké Google poskytuje při použití operátoru "define".
[2] Včetně mapového náhledu

5 Analýza základních vlastností vyhledávání

Jak bylo uvedeno již úvodu práce, cílem této a několika dalších kapitol není dokumentace všech funkcí vyhledávače Googlu, ani popis např. základních metod práce s operátory booleovské logiky. K těmto tématům již existuje velké množství literatury jak v klasických, tak elektronických zdrojích. Úkolem je analyzovat ne zcela zjevné principy fungování vyhledávacího mechanismu, popřípadě upozornit na vybrané pokročilé nástroje, užitečné z pohledu práce informačního specialisty.

5.1 Booleovská logika

5.1.1 Implementace booleovských operátorů v Googlu

Základní operace booleovské logiky, konjunkci, disjunkci a negaci, používají všechny velké vyhledávače (Google, Yahoo!, Live Search, Ask.com). Všechny též shodně používají konjunkci jako implicitní základní operaci, pokud u výrazu není uveden žádný operátor. Proto je v Googlu uvádění operátoru `AND` v dotazu nadbytečné (může však sloužit pro lepší přehlednost dotazu). Při konjunkci výrazů bere vyhledávač v úvahu jejich pořadí – dotaz `A AND B` dodá odlišně řazené výsledky než dotaz `B AND A`. Konjunkce v Googlu není tedy v tomto smyslu komutativní (narozdíl od disjunkce).

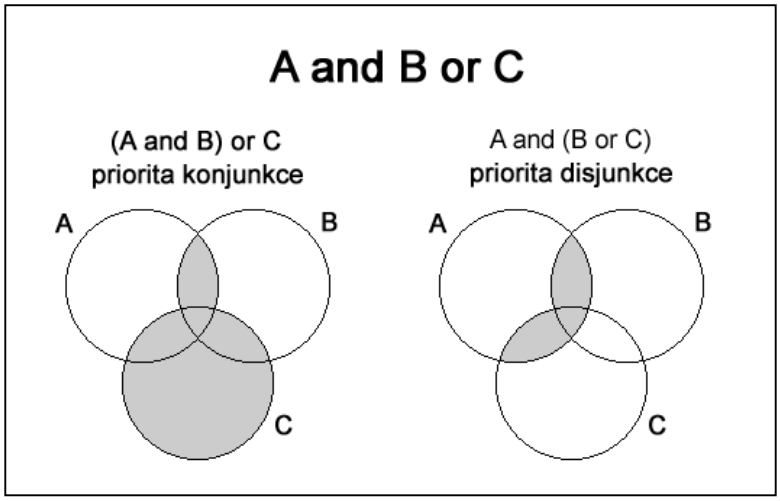
Disjunkce má v Googlu kromě standardního tvaru `OR` také alternativní podobu znaku `|` (kolmice). Operátor negace `NOT`, vyskytující se v některých vyhledávačích, Google nepodporuje, místo něj užívá jednoduchý znak `– (mínus)`.

Kromě těchto operátorů disponuje Google ještě operátorem `+`, který funguje jako striktnější podoba konjunkce. Pomocí něj lze vynutit vyhledávání takových výrazů, které by jinak Google z dotazu vypustil z důvodu jejich nízké informační hodnoty (tzv. „stop slova“). Operátor `+` lze použít také k vyhledání výrazu pouze v zadaném diakritickém tvaru (bez jeho dalších variant) a také zamezit hledání jeho odvozenin (viz kapitola 5.2.1).

Z pohledu základní teorie booleovské logiky není její implementace v Googlu úplná. Google totiž neumožňuje seskupování logických operací pomocí závorek (tzv. hnízdování, *nesting*), resp. nebere na případné závorky ohled.

Z tohoto hlediska se tak jeví jako důležitá informace, kterou ovšem Google nikde neuvádí – zda má při vyhodnocování složených dotazů větší prioritu konjunkce či disjunkce (a zda tedy odpovídá konjunktivní normální formě (KNF) či disjunktivní normální formě (DNF)).

Rozdíly mezi oběma možnostmi lze dobře ilustrovat s pomocí Vennových diagramů.



Obr. 5: Různá vyhodnocení stejného dotazu dle priorit operací

Pomocí vlastního dotazování bylo zjištěno, že Google patří do skupiny vyhledávačů, které uplatňují prioritu disjunkce. Notess dokládá [21], že i bez operátoru závorek lze v Googlu úspěšně pokládat složené booleovské dotazy. Samozřejmě s výjimkou těch, které vyjadřují prioritu konjunkce:

For this operation	Try this on Google
x AND y	`x y`
x OR y	`x OR y`
x AND (y OR z)	`x y OR z`
(x AND y) OR (z AND q)	**not yet possible**
(x OR y) AND (z OR q)	`x OR y z OR q`
x AND (y OR z OR q)	`x y OR z OR q`
x AND (y OR z) AND q	`x y OR z q`

Tab. 16: Možnosti složených dotazů. Zdroj: Searchengineshowdown.com

5.1.2 Srovnání s konkurencí

Některé vyhledávače upřednostňují na rozdíl od Googlu ve složených dotazech operaci konjunkce. Některé konkurenční stroje také podporují skládání více operací

pomocí závorek – lze v nich tedy specifikovat vlastní priority při vyhodnocování složených dotazů.

Vyhledávač	Prioritní operace	Nesting
Google	disjunkce	
Live Search	konjunkce	ano
Yahoo!	disjunkce	
Altavista	disjunkce	
Altavista Advanced Boolean Search	konjunkce	ano
AlltheWeb	disjunkce	
AlltheWeb Advanced Boolean Search	konjunkce	ano
Ask.com	konjunkce	

Tab. 17: Charakteristiky booleovských operací ve vyhledávačích

Mezi českými vyhledávači je z hlediska aplikace booleovské logiky výjimečný zejména Seznam, protože nepodporuje v žádné podobě operátory `AND`, `OR` či `NOT`. Konjunkce je prováděna standardně bez nutnosti operátoru, ostatní operátory nepodporuje z důvodu jejich nízké využívanosti [22].

Jiný způsob přístupu zvolil vyhledávač Centrum, který v rámci své vyhledávací technologie Morfeo umožňuje používat množství pokročilých operátorů [23]. Pomocí nich lze mimo jiné i přisuzovat jednotlivým výrazům v dotazu různou váhu, což je jinak jedním z výrazných omezení booleovské logiky ve vyhledávačích.

5.1.3 Nedostatky booleovského modelu pro vyhledávání dokumentů

Booleovskou logiku jako nástroj pro vyhledávání v množině indexovaných dokumentů používá většina internetových vyhledávačů. Tento model je velmi rozšířený z několika důvodů. Pro provozovatele vyhledávacích strojů je snadno implementovatelný a umožňuje rychlé vyhodnocování velkého množství dotazů, z pohledu uživatelů je logika snadno pochopitelná. Přesto má booleovský model ve své základní podobě několik omezení, jak je uvádí například Pokorný [24].

Efektivita a jednoduchost modelu je totiž důsledkem striktní interpretace booleovské logiky. Každá operace má tak pouze dva výsledné stavy – „pravda“ a „nepravda“ (true, false). To může někdy možnosti dotazu značně omezovat. Například dotaz `A AND B AND C AND D AND E` nevybere takové dokumenty, v nichž bude

chybět byť jediný z uvedených pěti výrazů. Pokorný uvádí, že takové dokumenty by nicméně také mohly uživatele zajímat a měly by být vybrány. Analogicky při dotazu `A OR B OR C OR D OR E` dokument obsahující pouze jeden výraz bude chápán jako stejně významný jako dokument obsahující všechny výrazy.

Stejně tak neexistuje v booleovské logice možnost, jak formulovat dotaz ve tvaru „vyhledej dokumenty, které obsahují aspoň 3 z celkových 6 zadaných výrazů".

Způsobem, jak zmírnit striktní interpretaci operátorů (Pokorný hovoří o jejich „změkčení") je změna modelu například na vektorový. Ten umožňuje pomocí skalárního součinu vektorů dokumentu a vektoru dotazu vyjádřit nejen jejich shodu, ale také míru této shody. Další možností je přiřazení vah k jednotlivým výrazům. Pomocí nich by bylo možné ohodnotit důležitost každého výrazu nejen pro daný dokument, ale i pro zadávaný dotaz.

Na straně dokumentů se tak již děje, algoritmy internetových vyhledávačů přisuzují jednotlivým výrazům v rámci indexovaného dokumentu různé váhy (podle jejich četnosti, pozice, přítomnosti v názvu dokumentu apod.). Přiřazování vah výrazům v dotazu naproti tomu podporuje jen zlomek vyhledávačů. Spíše než případné nároky na výpočetní techniku jsou důvodem nároky na uživatele – uživatelé obecně neradi tráví delší dobu tvorbou dotazu, když mají pocit, že stejně dobré výsledky získají i jednoduchým dotazem.

Existuje nicméně jedna velmi základní technika pro zvýšení váhy vybraného výrazu. Jedná se o tzv. *Google sinker*, jehož autorství je připisováno Philu Bradleymu [25]. Princip spočívá v pouhém několikanásobném opakování daného výrazu. Blakemanová uvádí příklad čtyř dotazů [26]:

```
beer market share belgium czech
beer market share belgium czech czech
beer market share belgium czech czech czech
beer market share belgium belgium czech czech czech
```

Pořadí, ale i počet nalezených výsledků se s postupným přidáváním mění, do popředí se dostávají dokumenty, které obsahují výraz `czech` vícekrát a jsou tedy v jistém smyslu k tomuto výrazu relevantnější. Jak Blakemanová uvádí, tuto techniku lze použít i v jiných vyhledávačích, např. v Yahoo!.

5.2 Stemming, lematizace, pravostranné rozšíření

5.2.1 Stemming

V listopadu roku 2003 pozměnil Google výrazným způsobem svůj vyhledávací algoritmus. Tato změna, internetovou komunitou nazývaná Florida Update, měla zásadní vliv na vyhodnocování dotazů. Kromě zadaných slov začaly být totiž vyhledávány také jejich další odvozené tvary – množná čísla u podstatných jmen, odvozená přídavná jména, různé rody a časy sloves apod. Pro toto odvozování používá Google termínu *stemming*, neboť je založené na vyhledávání všech slov, které mají stejný kmen (*stem*) jako výraz zadaný v dotazu.

Tuto podstatnou změnu ve vyhledávání Google (jak činí často) nijak oficiálně neprezentoval a nekomentoval. Jediná informace, kterou k této nové funkci poskytuje, je stručné vyjádření v jeho nápovědě [27]: „*Google now uses stemming technology. Thus, when appropriate, it will search not only for your search terms, but also for words that are similar to some or all of those terms.*“

Spolu s výhodou opětovného zvýšení uživatelského pohodlí (není třeba do dotazu zadávat různé tvary slov spojené disjunktním `OR`) přináší automatický stemming také řadu obtíží.

Ne vždy musí být automatické hledání odvozených tvarů uživatelem žádané. Stemming však nelze nikde centrálně aktivovat či deaktivovat. Je možné pouze každý jednotlivý výraz v dotazu uzavřít do uvozovek (popř. použít operátor `+`), čímž se zabrání jeho odvozování.

Především však není vůbec zřejmé, které výrazy z dotazu Google podrobil stemmingu a které nikoli. Google tuto informaci při výpisu výsledků neposkytuje (narozdíl např. od českého Jyxa). Některé starší zdroje uvádějí [28; 29], že Google nepoužívá stemming pro jednoslovné dotazy. Vlastním pokusným dotazováním však bylo zjištěno, že v současné době hrají odvozeniny roli už i v případě některých jednoslovných dotazů – například u dotazů `window`, `fires` či `dropping`. Pomocí metodiky použité a popsané v kapitole 4 byla sledována změna ve složení prvních třiceti výsledků při vyloučení stemmingu pomocí operátoru `+`. Změna je vyjádřena ve formě shodnosti obou sad výsledků.

Dotaz	Shodnost výsledků	Dotaz s vyloučením stemmingu
window	57 %	+window
fires	80 %	+fires
dropping	63 %	+dropping

Tab. 18: Vliv stemmingu na vybrané jednoslovné dotazy

V případě dotazu window figuruje ve výsledcích především velké množství dokumentů jen s odvozeným tvarem windows, přítomným v názvu stejnojmenného operačního systému. Algoritmus při vyhledávání v indexu sice přisuzuje větší váhu dokumentům, v kterých se nachází výraz přímo v zadaném tvaru, než dokumentům, v kterých se vyskytují jen jeho odvozeniny. Nicméně lze předpokládat, že vlivem velkého množství dalších faktorů, které algoritmus bere v potaz, se nakonec mohou umístit ve výsledcích výše i dokumenty jen s těmito odvozeninami.

Další problém s předchozím přímo souvisí – není jasné, jaké všechny tvary stemmer Googlu od zadaného výrazu vytvořil a které jsou tedy vyhledány. Vodítkem je opět pouze vizuální kontrola textu vypsaných výsledků, odvozené tvary jsou v něm zvýrazněny tučným písmem.

Situaci dále komplikuje fakt, že funkci stemmeru ovlivňuje také jazyková verze uživatelského rozhraní. Například zmíněný dotaz window, pokud je zadán v českém rozhraní, vyhledá pouze tento jeden tvar (na rozdíl od rozhraní anglického). A naopak, pouze v českém rozhraní se ve výsledcích dotazů začínají objevovat také odvozené tvary českých slov. Při dotazu parkoviště +náměstí je tak v textu výsledků zvýrazněno také slovo parkování; v případě druhého dotazu má vliv na výsledky také přítomnost slova místa.

Dotaz	Shoda	Dotaz s vyloučením stemmingu
parkoviště +náměstí	60 %	+parkoviště +náměstí
místo +pracovní	37 %	+místo +pracovní

Tab. 19: Vliv stemmingu na vybrané dotazy v českém jazyce

5.2.2 Srovnání s konkurencí

Zatímco na poli globálních internetových vyhledávačů je Google se svým stemmingem spíše výjimkou, v českých podmínkách je hledání ohýbaných tvarů slov již delší dobu důležitou funkcí tuzemských vyhledávacích služeb. Čeština má narozdíl od angličtiny mnohem ohebnější tvarosloví – skloňováním podstatných jmen a časováním sloves vzniká velké množství různých tvarů téhož slova, často se též mění i samotný jeho kmen.

Pro vyhledávání různých tvarů slov se v češtině používá proces lematizace (a následné derivace). Ač funguje na podobném principu jako stemming, není to termín totožný a zaměnitelný [30].

Lematizaci používají všechny tři velké české portály:

- Seznam používal v letech 2003 až 2005 služeb vyhledávače Jyxo, který pro lematizaci používá vlastní slovník Ispell vyvíjený pod licencí General Public License (GPL). Od jara 2005, kdy Seznam začal používat vlastní fulltextové vyhledávání, používá pro lematizaci Český lematizátor od společnosti Lingea [31]. Hledání ohýbaných slov nelze vypnout.

- Centrum využívá služeb technologie Morfeo, která používá vlastní lematizátor. Ohýbání lze aktivovat a deaktivovat v sekci Rozšířené hledání.

- Atlasu poskytuje vyhledávací služby vyhledávač Jyxo se svým slovníkem Ispell. Při hledání na Atlasu nelze ohýbání vypnout, ovšem pokud uživatel použije přímo rozhraní Jyxa[1], je mu kromě tohoto nastavení přístupný i seznam všech variant slov, která při zapnuté lematizaci Jyxo k zadanému slovu vyhledá.

5.2.3 Pravostranné rozšíření (truncation)

Pravostranné rozšíření je další možností jak specifikovat v dotazu varianty slov, které se mají vyhledávat. Tato funkce je důležitou a běžnou součástí rešeršního procesu například v dialogových rozhraních profesionálních databázových center. Ve sféře internetových vyhledávačů s globální působností a především globální přístupností je to však funkce velmi ojedinělá. Důvodem proti volnému zavedení tohoto operátoru

[1] http://www.jyxo.cz

ve větším měřítku je nebezpečí nekontrolovatelného nárůstu nároků na výpočetní kapacity. Jediné slovo s necitlivě použitým pravostranným rozšířením (např. `dis*`) by tak znamenalo vyhledávání až mnoha desítek variant, které přitom spolu nemusejí sémanticky vůbec souviset. Ještě více by výkonnost stroje poznamenalo hypotetické levostranné rozšíření, neboť by musel existovat zvláštní index všech slov v reverzním tvaru.

Nebezpečí nezvladatelného náporu na hardware je tak jedním z důvodů, proč vyhledávače obecně nenabízejí příliš pokročilé, na výpočetní čas náročné operátory. V poslední době byl například Live Search donucen pozastavit funkčnost některých svých pokročilejších operátorů – systém byl záměrně pomocí nich zatěžován, pravděpodobně za účelem hromadného vytěžování dat [32].

Lze také předpokládat, že Google nepovažuje pravostranné rozšíření za výrazný přínos pro uživatele. Typický uživatel chce co nejrychleji získat relevantní výsledky bez toho, aby k jejich dosažení musel používat pokročilé operátory. Tímto směrem také Google svůj systém zdokonaluje – funkčnost pravostranného rozšíření nahrazuje stále rozvinutějším stemmingem, nabídkou synonym, opravou překlepů apod.

Některé vyhledávače v minulosti pravostranné rozšíření umožňovaly, ale upustily od něj (Altavista). Pravděpodobně jediným větším vyhledávačem, který pravostranné rozšíření v současné době podporuje, je původem francouzský systém Exalead.[1]

Z českých vyhledávačů disponuje operátorem pro pravostranné rozšíření jen Morfeo, které však jeho použití (pravděpodobně z výše uvedených důvodů) omezuje pouze na slova s minimální délkou tři znaky [23].

5.3 Vzdálenostní operátory

5.3.1 Rozdělení vzdálenostních operátorů

Jako součást nebo doplněk booleovské logiky bývají občas uváděny vzdálenostní (proximitní) operátory, které při zadání více výrazů blíže specifikují jejich vzájemnou pozici v hledaném dokumentu.

[1] http://www.exalead.com

Tyto operátory se dají rozdělit zhruba do tří skupin:

- Operátor určující bezprostřední blízkost výrazů – ty se tak nacházejí v textu přímo vedle sebe, neleží mezi nimi již žádná další slova. V naprosté většině vyhledávačů se k tomuto účelu používá operátor uvozovek `""`, kterým se z obou výrazů vytvoří *fráze*. Alternativním řešením je v některých vyhledávačích používání operátorů, které narozdíl od fráze vyhledají oba sousedící výrazy v libovolném pořadí (např. operátor `NEXT` ve vyhledávači Exalead).

- Operátor jako zástupný znak, „divoká karta" (wildcard), nahrazující v textu libovolné jedno slovo. Pomocí opakovaného použití tohoto operátoru (nejčastěji v podobě `*`, tzv. hvězdičkové notace) mezi výrazy v dotazu lze vyjádřit jejich různou vzdálenost. Jedná se o operátor typicky používaný při tvorbě pravostranného rozšíření (či regulárních výrazů), v tomto případě ovšem nezastupuje znaky, ale celá slova.

- Číselný operátor, určující maximální vzdálenost výrazů v textu (např. ve tvaru `alfa n5 beta`). Narozdíl od zástupných znaků definuje vzdálenost výrazů nikoli jednou diskrétní hodnotou, ale pomocí celého intervalu (např. „nejvýše pět slov od sebe"). Podle konkrétní implementace může a nemusí hrát roli zadané pořadí výrazů. V některých systémech může být také velikost intervalu omezena na jednu pevně stanovenou hodnotu – nejčastěji zhruba na 10 slov.

Google používá jen dva ze tří výše uvedených operátorů: frázi a zástupný operátor (wildcard).

Nicméně vzdálenost výrazů v textu hraje v Googlu roli vždy, i v případech, kdy není použit ani jeden z nich. Čím menší je vzájemná vzdálenost výrazů, tím vyšší váha je danému dokumentu přisouzena v procesu konečného pořádání výsledků. Jelikož ale algoritmus Googlu bere v úvahu velké množství dalších kritérií (viz kapitola 2.3), nemusí se faktor blízkosti či vzdálenosti výrazů v konečném pořadí viditelně projevit.

5.3.2 Operátor fráze

Frází je nazývána posloupnost dvou či více slov, která mají být vyhledána pouze v této posloupnosti. Použití horních uvozovek pro tvorbu fráze je v současných vyhledávačích nepsaným standardem. V Googlu lze pomocí fráze také zabránit vyhledávání odvozených tvarů (stemming).

Z pohledu českého uživatele má však operátor fráze také jednu vedlejší, skrytou vlastnost, která souvisí s diakritikou. Pokud je jednoslovný výraz uzavřen do uvozovek (a tedy zapsán jako fráze), je vyhledán pouze v zadaném diakritickém tvaru. To ovšem neplatí, pokud se fráze skládá z více slov (rozhodující je už samotná přítomnost mezery). V takovém případě je nutné dodatečně uvnitř fráze použít operátor +, jinak budou vyhledávány i ostatní diakritické tvary všech slov ve frázi. Konkrétní chování vyhledávače navíc závisí také na zvoleném jazyce uživatelského rozhraní. Problematika diakritiky u jednoslovných výrazů je podrobněji popsána v kapitole 4.3.

<table>
<tr><th>dotaz</th><th></th><th colspan="2">anglické rozhraní</th><th>české rozhraní</th></tr>
<tr><td>Dvorak</td><td rowspan="9">vyhledá dokumenty obsahující diakritické varianty</td><td colspan="3">různé</td></tr>
<tr><td>Dvořák</td><td colspan="2">různé</td><td>jen zadanou</td></tr>
<tr><td>“Dvorak“</td><td colspan="3">jen zadanou</td></tr>
<tr><td>“Dvořák“</td><td colspan="3">jen zadanou</td></tr>
<tr><td>“Antonin Dvorak“</td><td colspan="3">různé</td></tr>
<tr><td>“Antonín Dvořák“</td><td colspan="2">různé</td><td>jen zadanou</td></tr>
<tr><td>“+Antonin +Dvorak“</td><td colspan="3">jen zadanou</td></tr>
<tr><td>“+Antonín +Dvořák“</td><td colspan="3">jen zadanou</td></tr>
</table>

Tab. 20: Role diakritiky, operátoru fráze a operátoru + v jednoslovných a víceslovných dotazech

5.3.3 Zástupné znaky (wildcards)

Jediným dalším vzdálenostním operátorem, který Google podporuje, je zástupný znak * (hvězdičková notace). Tento operátor je v systému přítomen již od roku 2002, ale dlouho nebyl nikde oficiálně zdokumentován a jeho funkčnost byla dlouho nejasná [33]. Teprve v roce 2005 byl popsán v oficiálním blogu Googlu [34] a uveden také v nápovědě vyhledávače [35]:

*„A wildcard in a Google search query can be indicated by an asterisk and will match one or more entire words of text so that the query matches a contiguous sequence of words. For example, a search for [cooking * classes] will match the phrases "cooking school classes" and "cooking and wine tasting classes." ... Also, more than one asterisk can be used, such as [vitamin * is good for *].“*

Tento obecný popis opět přesnou funkci zástupného operátoru příliš nevymezuje. V různých dotazech tak jeden výskyt znaku * představuje jedno, dvě nebo i více slov. Google nicméně dává najevo, že účelem tohoto operátoru není zastupovat pevně stanovený počet slov, ale spíše obecně nahrazovat libovolně dlouhý úsek textu mezi dvěma zadanými výrazy. Nazývá to „doplňováním prázdných míst“ (filling in the blanks). Notess uvádí [36] několik dalších možností tohoto využití hvězdičkové notace, např. při odhalování sofistikovanějších případů plagiátorství, či hledání variantních reklamních sloganů.

V situaci, kdy není pevně stanoveno, kolik slov jeden znak * zastupuje, je o to více nejistá funkce několika těchto znaků zřetězených za sebou, např. "A * * * B". Je zřejmé, že uvedený příklad nevyhledá pouze takové dokumenty, ve kterých se vyskytují výrazy A a B *přesně* tři slova od sebe. Stejně tak ale nefunguje ani jako standardní číselný proximitní operátor, který by vyhledal výrazy A a B ve vzdálenosti jednoho, dvou či tří slov, navíc v jakémkoliv pořadí.

Někteří autoři se proto zabývali vývojem algoritmu, který by přesněji vymezil a tím i zdokonalil funkci tohoto operátoru v Googlu.

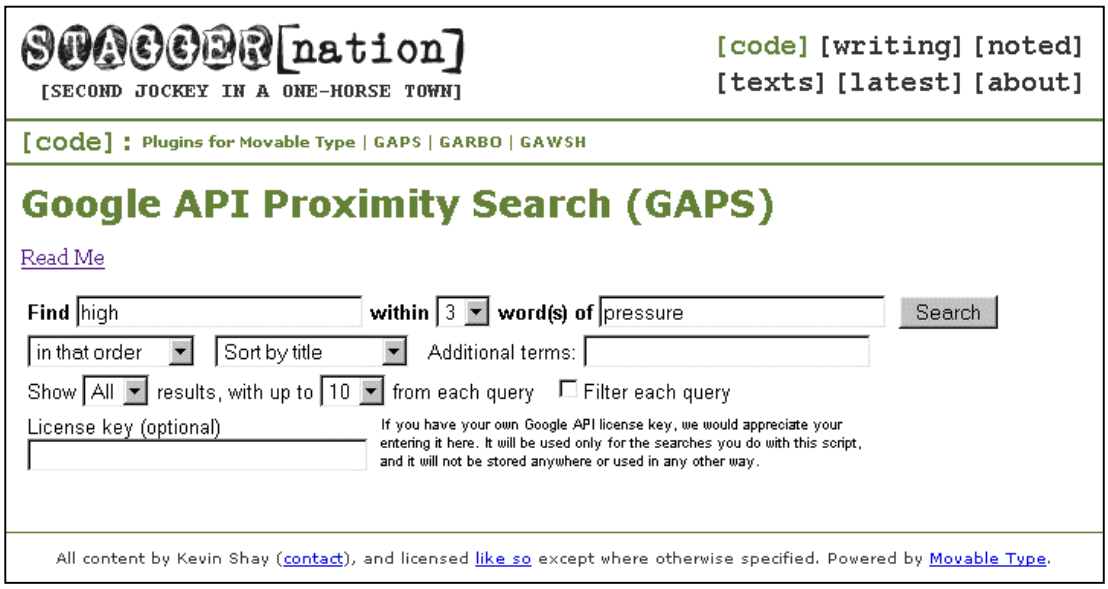
Kostoff [37] se snažil vymezit počet zastupovaných slov omezením seshora. Každý dotaz s definovanou vzdáleností X tedy doplnil o negaci stejných výrazů ve vzdálenosti X+1, tedy například "A * * B" -"A * * * B".

Myšlence přiblížit googlovský zástupný znak co nejvíce klasické proximitní funkci NEAR, známé z profesionálních rešeršních služeb, se věnoval Kevin Shay. Vytvořil veřejně přístupnou aplikaci GAPS (Google API Proximity Search) [38], která pomocí formulářového rozhraní umožňuje vyhledávání dvou výrazů ve vzájemné vzdálenosti až tří slov. Děje se tak prostřednictvím několika dílčích dotazů, jejichž počet závisí na zvolené vzdálenosti.

Při maximální vzdálenosti tří slov se tak v případě výrazů A a B provede těchto osm dotazů:

```
"A B"                    "A * * B"
"B A"                    "B * * A"
"A * B"                  "A * * * B"
"B * A"                  "B * * * A"
```

Prvních deset výsledků od každého dotazu je následně sloučeno do jednoho výpisu. Ten je seřazen podle zvoleného kritéria (název, URL, vzdálenost výrazů, pořadí výsledku v rámci dílčího dotazu).



Obr. 6: Google API Proximity Search. Zdroj: Staggernation.com

Maximální nastavitelná vzdálenost tří slov je zvolena s ohledem na počet dílčích poddotazů, pokládaných Googlu. Jejich počet při vzdálenosti X činí 2*(X+1). Každá aplikace, vytvořená s pomocí Googlem poskytovaného programátorského rozhraní pro tvorbu aplikací (API), má stanoven limit – tisíc položených dotazů za den. Aby mohla být aplikace veřejně použitelná, stanovil proto autor výše zmíněnou omezující podmínku.

5.3.4 Srovnání s konkurencí

Vedle standardního operátoru fráze nejsou ve vyhledávačích pokročilejší formy proximitního vyhledávání příliš rozšířené. To platí nejen pro zástupné znaky, ale především pro číselné operátory – tato pokročilá varianta vzdálenostního operátoru je sice běžná pro vyhledávání v profesionálních rešeršních systémech, ale v prostředí internetových vyhledávačů se takřka nevyskytuje. V minulosti umožňovala takovéto vyhledávání Altavista, po převzetí tohoto vyhledávače konkurenční firmou Yahoo! však byla tato funkce (stejně jako pravostranné rozšíření) ukončena.

Hvězdičkovou notaci lze kromě Googlu použít ještě v Yahoo!, jde však o nedokumentovanou funkci, která už byla několikrát zavedena a pozastavena [39].

Nejblíže se pojetí klasického číselného proximitního operátoru přibližuje (v oblasti velkých světových vyhledávačů) jen Exalead, který používá operátor `NEAR`. Podle nápovědy k vyhledávání [40] není hodnota operátoru nastavitelná, je pevně stanovena na maximální vzdálenost 16 slov. V březnu roku 2006 bylo však zjištěno a následným testováním potvrzeno [41], že operátor `NEAR` lze používat i ve tvaru `A NEAR/X B`, kde X je požadovaná vzdálenost obou výrazů.

Dalším proximitním operátorem vyhledávače Exalead je `NEXT`, který plní stejnou funkci jako operátor fráze, ale bez ohledu na pořadí výrazů. Jde tedy vlastně o ekvivalent dotazu `"A B" OR "B A"`.

České vyhledávače nabízejí v oblasti vyhledávání podle vzdálenosti jen možnost vytvořit z výrazů frázi. Jedinou výjimkou je Centrum (resp. Morfeo), které nabízí stejně jako Google možnost použít ve frázi hvězdičkovou notaci pro reprezentaci celých slov. Jaký má však zástupný znak `*` přesný rozsah v nápovědě neuvádí [23]. Vlastním pokusným dotazováním bylo zjištěno, že se jedná o rozsah malý, v řádu jednoho či dvou slov. Operátor navíc nelze řetězit.

6 Specifika rešeršní činnosti v prostředí Googlu

6.1 Řízené slovníky

Základním prvkem, důležitým pro vyhledávání v klasických („neinternetových“) dokumentografických informačních systémech, je určitá forma řízeného slovníku (tezaurus, předmětový heslář, desetinné třídění aj.). Při vstupním zpracování přiřadí katalogizátor každému dokumentu skupinu popisných dat (metadat), které tvoří bibliografický záznam dokumentu. V něm je dokument také tematicky a obsahově klasifikován právě výběrem prvků řízeného slovníku (deskriptorů, hesel). Pomocí této klasifikace je tak možné vyhledat nejen daný dokument, ale například také velmi přesně určit všechny další, tematicky příbuzné.

Právě vysoká strukturovanost bibliografického záznamu, umožňující precizní rešeršní činnost, je hlavní výhodou oproti internetovým vyhledávačům, jako zástupcům systémů založených především na volnotextovém vyhledávání (viz dále).

Oproti internetovému vyhledávání má ovšem vyhledávání s pomocí řízených slovníků také několik nevýhod:

- Klade větší nároky na poskytovatele vyhledávání – je nutné slovník vytvořit, resp. udržovat.
- Klade větší nároky na zpracování záznamu dokumentu – je nutné zabezpečit kvalifikované katalogizátory.
- Klade větší nároky také na uživatele – formulace dotazu vyžaduje určité větší než základní znalosti.

6.2 Volnotextové vyhledávání

Internetové vyhledávače jako Google byly vždy vyvíjeny se zřetelem na charakteristické znaky množiny cílových dokumentů – webových stránek. Tato množina je specifická především dvěma extrémy:

1. svou velikostí – jen samotný Google v současné době pokrývá podle některých odhadů [42] řádově až desítky miliard dokumentů.

2. svou různorodostí – především ve smyslu rozličných formátů dokumentů, jejich struktur, jazyků a různé míry dostupných metadat

Zatímco z hlediska extrémní různorodosti by byla přítomnost člověka v procesu zpracování velmi žádoucí a nezastupitelná (neboť zcela automatizovaný katalogizační proces není za těchto podmínek možný), extrémní množství dokumentů lidský faktor prakticky vylučuje.

Všechny globální internetové vyhledávače, pokrývající tak rozsáhlou množinu dokumentů, musejí být nutně založeny čistě na strojovém zpracování. Tento způsob zpracování má několik podstatných důsledků:

- Bez přítomnosti člověkem prováděné analýzy je katalogizační proces v podstatě zredukován pouze na indexaci kompletního textového obsahu dokumentu (s výjimkou tzv. stop slov, které některé systémy neindexují).

- Kromě svého obsahu je dokument v systému popsán již jen minimem metadat – adresou URL, datem poslední indexace a seznamem stránek, které na dokument odkazují.

- Indexace plného textu dokumentu se může na první pohled jevit jako ideální z hlediska definice jeho obsahu. Je to řešením jedné velké nevýhody systémů založených na řízených slovnících, v nichž deskriptory či hesla přiřazené dokumentům nemohou plně postihovat všechny nuance jejich obsahu – vždy se totiž jedná o určité zobecnění. Tato výhoda internetových vyhledávačů však v prostředí miliard indexovaných dokumentů přechází paradoxně až v nevýhodu. Pouhé plné znění dokumentu nemusí stačit pro vystihnutí a charakteristiku jeho obsahu. Výrazně zde chybí přítomnost určité formy zobecněného popisu v podobě sémantických metadat, jak je nazývá např. Sklenák [43], jako vyšší stupeň nad metadaty syntaktickými a strukturálními.

- Protože dokumenty jsou v systému reprezentovány nikoli prostřednictvím prvků řízeného slovníku, ale především svými plnými texty, lze v nich vyhledávat pomocí výrazů přirozeného jazyka – tzv. volnotextovým vyhledáváním (free-text search).

- Volnotextové vyhledávání nabízí velkou volnost při výběru klíčových slov a je proto považováno za uživatelsky přívětivé a nenáročné.

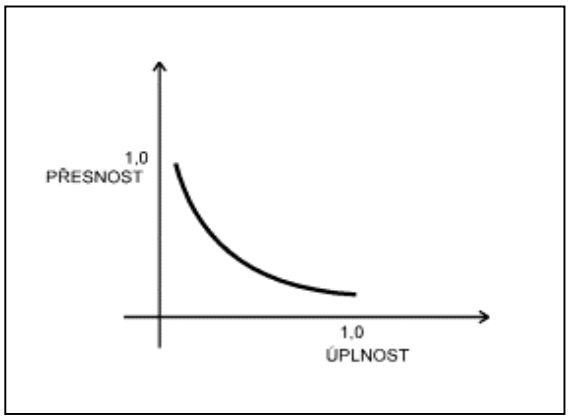
- Hledané dokumenty je možné definovat jen s pomocí klíčových slov z jejich obsahu, což neumožňuje takovou přesnost vyhledávání jako definice pomocí metadat a řízených slovníků. Proto je nutné klíčová slova vybírat (a kombinovat) velmi pečlivě, jinak dochází k vyhledání velkého množství nerelevantních dokumentů (viz následující kapitola).

- Zdánlivá jednoduchost tvorby dotazu a velké množství (byť méně relevantních) výsledků na obecné dotazy ale uživatele ke zmíněné pečlivosti nijak nemotivuje.

6.3 Vyhledávání z hlediska koeficientů úplnosti a přesnosti

Efektivita vyhledávacích strojů je definována především jako míra schopnosti vyhledat relevantní dokumenty. Tato schopnost je vyjadřována pomocí koeficientů **úplnosti** a **přesnosti** vyhledávání. Vzájemný vztah hodnot obou koeficientů vymezuje, s jakými problémy se vyhledávací stroj z hlediska relevance potýká. Pro Google a ostatní internetové vyhledávače je (narozdíl od klasických rešeršních systémů) charakteristická relativně vysoká míra koeficientu úplnosti **R** a naopak nízká hodnota koeficientu přesnosti **P.**

Ideální hodnota obou koeficientů je co nejvyšší, takové vyhledávání by tedy nalezlo všechny relevantní dokumenty v systému a žádné nerelevantní. To není nicméně reálně možné, neboť mezi přesností a úplností panuje vztah nepřímé úměrnosti, čím vyšší úplnost, tím nižší přesnost a naopak.



Obr. 7: Hrubé znázornění vztahu parametrů přesnosti a úplnosti [44]

6.3.1 Koeficient úplnosti R (recall)

- Udává poměr nalezených relevantních dokumentů a všech relevantních dokumentů v systému.
- Vyjadřuje pravděpodobnost toho, že relevantní záznam bude vybrán.
- Při nízké hodnotě úplnosti bude vyhledána jen malá část všech relevantních dokumentů.

$$\mathbf{R} = \frac{\text{počet vyhledaných relevantních dokumentů}}{\text{celkový počet všech relevantních dokumentů v systému}}$$

Vzhledem ke zcela automatickému zpracování plných textů a z toho plynoucí možnosti volnotextového vyhledávání lze říci, že míra úplnosti vyhledávání je v Googlu relativně vysoká. To platí především pro srovnání se systémy s vyhledáváním pomocí řízených slovníků. V nich nemusí být nalezeny relevantní dokumenty v případech, kdy je například hledaný výraz příliš specifický či okrajový, takže v řízeném slovníku nefiguruje jako deskriptor.

6.3.2 Koeficient přesnosti P (precision)

- Udává poměr nalezených relevantních dokumentů a všech nalezených dokumentů.
- Vyjadřuje pravděpodobnost toho, že vybraný záznam bude relevantní.
- Při nízké přesnosti je vyhledáno také mnoho irelevantních dokumentů.

$$\mathbf{P} = \frac{\text{počet vyhledaných relevantních dokumentů}}{\text{celkový počet vyhledaných dokumentů}}$$

Naopak právě v nízké míře přesnosti je základní slabina Googlu. Podstatu vyhledávání v Googlu tvoří takřka výhradně procházení plných textů dokumentů. Proto je společně s relevantními dokumenty vyhledáno také velké množství

dokumentů, v nichž se hledaný výraz sice fakticky nachází, ale v jiném (k dotazu nerelevantním) kontextu. To platí především pro obecně formulované dotazy.

Na rozdíl od Googlu je v systémech založených na řízených slovnících poměr hodnot obou koeficientů opačný – relativně nižší úplnost, ale vyšší přesnost. To je důsledkem nemalého vkladu lidské práce, sémantická analýza obsahu při zpracování dokumentů zde hraje výraznou roli.

6.3.3 Koeficient přesnosti ve volnotextových vyhledávačích

Nevýhody nízké přesnosti volnotextového vyhledávání lze řešit různými způsoby.

Především vytvářením komplexnějších dotazů, zapojením pokročilých funkcí, pokud je vyhledávač nabízí. Tedy kombinováním výrazů booleovské logiky, užíváním frází, proximitního vyhledávání, regulárních výrazů apod.

Volnotextové vyhledávání lze také zpřesnit tím, že se omezí jen na určité oblasti dokumentu. V případě Googlu tak lze učinit prostřednictvím operátorů `intitle:`, `inurl:`, `inanchor:` a `intext:` (viz kapitola 7.1).

Sám autor může také charakterizovat obsah dokumentu pomocí relevantních klíčových slov, která do něj vloží. To je obzvláště užitečné, pokud se tato slova v textu vůbec neobjevují. Lze je specifikovat (v případě webových stránek) v hlavičce zdrojového kódu jako hodnotu značky (tagu) `<meta name="keywords">`. V minulosti se takto zadanými klíčovými slovy vyhledávače při zpracování dokumentu řídily, bohužel s nástupem intenzivního používání (a zneužívání) SEO došlo k naprosté devalvaci tohoto tagu. Dnes už vyhledávače jeho hodnotu pomíjejí.

Provozovatelé internetových vyhledávacích strojů se snaží nízkou přesnost kompenzovat tím, že často odkazované, a tedy pravděpodobně hodnotnější dokumenty posunují ve výsledcích na vyšší pozice. Jedná se však jen o pomocný prvek, citovanost nesouvisí s relevancí vůči uživatelově dotazu.

Nejsystematičtějším řešením by bylo důsledné používání některého metadatového standardu (například Dublin Core). Tento posun směrem k „sémantickému webu“ by umožnil precizní strukturalizaci dokumentu. I toto řešení má však svá negativa – jeho

předpokladem je větší rozšíření jazyka XML, resp. XHTML jako standardu pro tvorbu webových stránek. Navíc i ve strukturovaném dokumentu by mohla metadata přinášet záměrně zavádějící informace – v důsledku zneužívání metod SEO, tak jak se tomu stalo již u zmíněného tagu `<meta name="keywords">`.

6.4 Definice a role relevance

Pojem relevance bývá nejčastěji definován jako míra významnosti pro určité téma, jako věcná příslušnost k položenému dotazu. Podrobněji jsou odlišovány tři přesnější druhy relevancí, jak je uvádí například Kučerová [45]:

- **formální** relevance – relevance na úrovni syntaxe. Vyjadřuje základní faktickou shodu textu dotazu s textovým obsahu dokumentu. Tuto relevanci určuje vyhledávací program.

- **věcná** relevance – relevance na úrovni sémantiky. Vyjadřuje shodu významové blízkosti mezi obsahem dokumentu a položeným dotazem.

- **pertinence** – subjektivní relevance. Vyjadřuje vztah mezi obsahem dokumentu a subjektivní informační potřebou uživatele. Je značně proměnlivá v čase.

Všem zcela automatizovaným vyhledávacím systémům je vlastní vyhledávání dokumentů pouze na nejnižší úrovni, na bázi formální relevance. Posun k relevanci věcné by byl u těchto systémů možný jen s pomocí nějaké pokročilé formy sémantické analýzy či umělé inteligence založené na hierarchických řízených slovnících. To by umožnilo stroji například pochopit souvislost mezi slovy „déšť“ a „pršet“.

Současné internetové vyhledávače však těmito mechanismy nedisponují [46]. Jazyková umělá inteligence, je li možné to tak nazvat, se v Googlu omezuje pouze na netransparentně fungující stemming (viz kapitola 5.2.1), opravu pravopisných překlepů a velmi základní formu slovníku synonymních výrazů (kapitola 7.1.1).

Jádro vyhodnocování dotazu však stále pracuje především na úrovni formální relevance, na prosté shodě kódu dotazu s kódem dokumentu. Google tak v podstatě klade rovnítko mezi výrazy, kterými se uživatel snaží identifikovat hledaný dokument a mezi výrazy, které jsou obsaženy v textu jakéhokoliv dokumentu.

Je tedy důležité, aby uživatel identifikoval dokument nikoliv intuitivně, pomocí výrazů definujících jeho smysl, ale skutečně jen pomocí slov, která se mohou vyskytovat v jeho obsahu. V případě prvního, intuitivního způsobu tvorby dotazu lze očekávat opět nižší míru přesnosti.

6.4.1 Důsledky formálně relevantního vyhledávání pro rešeršní činnost

Relevance založená na čisté shodě kódu dotazu s kódem dokumentu může mít vliv nejen na nízkou přesnost, ale také na nízký koeficient úplnosti. Některé relevantní dokumenty mohou tak zůstat nevyhledány. To se může týkat kupříkladu dotazů obsahujících určitý druh obecného geografického vymezení. Dotaz hledající informace o důlních neštěstích v Austrálii tak nenalezne takové dokumenty, v nichž se slovo „Australia“ ani „australian“ nemusí teoreticky ani jednou vyskytovat – v textu může být popisováno jen konkrétní místo (např. Melbourne). Určitým řešením zde může být omezení hledání Googlu pouze na australské stránky, prostřednictvím sekce Jazykové nástroje. Tento způsob však nelze použít vždy – například pokud by se jednalo o hledání informací o lidových slavnostech ve východních Čechách.

Dalším případem, kdy nemusí být snadné určit hledaný dokument jen pomocí slov v něm obsažených je časové vymezení. Dokumenty popisující například „minulý vývoj“, „současný stav“ či „prognózu budoucího vývoje“ v libovolném oboru lidské činnosti nemusí tato slova v textu vůbec obsahovat.

Jiným problémem z hlediska úplnosti výsledků hledání je používání negace v dotazu. Pokud bude uživatel hledat kupříkladu informace o různých poruchách učení, ovšem s výjimkou dyslexie, pravděpodobně ji z dotazu vyloučí pomocí operátoru negace: `“-dyslexie“`. Tím však z výsledků vyřadí i všechny takové dokumenty, které mohou být cele zaměřeny na jiné poruchy učení, ale termín „dyslexie“ jen okrajově zmiňují.

Zcela zavádějící výsledky může mít vyhledávání, při kterém se uživatel snaží definovat žádané informace podle jejich abstraktních vlastností. Tedy nejen podle jejich obsahu, ale také např. pomocí subjektivního hodnocení – ve formě adjektiv „přesný“, „pravdivý“, „odborný“, „kvalitní“, „levný“ apod. Na takovéto přívlastky (zejména na poslední uvedené) se totiž pomocí metod SEO specializují weby

komerčních firem či internetových obchodů, výsledky dotazů jsou tímto faktem pak značně ovlivněny. Obecně je vyhledávání takto subjektivně vymezených informací jen prostřednictvím vyhledávačů problematické, ne-li nemožné.

6.4.2 Jazykové nároky na rešeršní činnost z hlediska formální relevance

Z podstaty formální relevance uplatňované v Googlu vyplývá velká důležitost každého jednoho znaku tvořícího dotaz. Z tohoto důvodu je nutné, aby uživatel (a obzvláště informační specialista) měl také odpovídající jazykovou průpravu, která je nezbytným předpokladem pro efektivní práci s českými i cizojazyčnými výrazy vkládanými do dotazu.

Relativně malý problém představují případy polysémie či homonymie některých českých pojmů, kdy jsou vyhledány navíc také jejich nevyhovující významy. Jsou totiž ve výsledcích jasně patrné a lze je vyloučit buď přidáním dalších zpřesňujících výrazů (např. u různých významů slova koruna apod.), popřípadě použitím operátorů + a **“ “** u nepravých homonym (např. plast, plášť, plást).

Závažnějším a méně zjevným problémem jsou naopak případy, kdy ve výsledcích některé relevantní dokumenty chybí. Slovo zadané v dotazu se totiž může na internetu vyskytovat také v dalších tvarech, z formálního hlediska více či méně správných. Odchylky těchto tvarů mohou mít různé podoby:

- Jednotlivé tvary se mohou lišit jen v zápisu diakritiky – např. dvojice slov balon/balón, stadion/stadión, sezona/sezóna, masivní/masívní aj.

- Rozdíly v jediném znaku jsou typické zejména pro české přepisy cizích slov – např. koncovky ve slovech altruismus/altruizmus, marketink/marketing, také kurs/kurz, diskuse/diskuze aj.

- Větší rozdíly se mohou objevovat u slov, u nichž se neujal výhradně jeden tvar, popřípadě se používají často i patvary, nebo i zcela gramaticky chybné varianty – např. webový/webovský, dehydrace/dehydratace, potenciální/potencionální, standardní/standartní. V posledním jmenovaném případě je však již na místě zvažovat hodnotu dotyčného dokumentu.

V pokročilých rešeršních systémech by k vyhledání různých tvarů téhož pojmu bylo použito zástupných znaků, popřípadě pravostranného rozšíření, v Googlu je nutné použít disjunkci všech pravděpodobných variant.

Pro vyhledávání v češtině může být vhodné do dotazu začlenit také počeštěné varianty cizích slov, například intravenózní/nitrožilní, management/řízení apod.

Odpovídající jazykové dovednosti jsou důležité také pro vyhledávání v cizojazyčném jazykovém prostoru.

Google pro anglický jazyk uplatňuje při vyhledávání lematizaci (resp. stemming). Ta je užitečná pro automatické vyhledávání různých odvozenin zadaných výrazů. Rešeršér by však kvůli netransparentnímu chování neměl na stemmer zcela spoléhat a opět použít jistější disjunkci odvozených tvarů. Totéž lze říci o rozdílných zápisech slov v britské a americké angličtině – např. u variant behavior/behaviour, swiming/swimming, ale i sidewalk/pavement apod.

Při hledání napříč jazyky je nutné zvážit také přepisy zdánlivých internacionalizmů v názvech, jako jsou např. USA (ve francouzštině „EU“) nebo transkripce osobních jmen (Gorbachov/Gorbachev/Gorbatschow/Gorbatchev aj.).

6.4.3 Relevance, úplnost a přesnost ve vztahu k rešeršní strategii

Úplnost a přesnost vyhledávání má také vztah k použité rešeršní strategii. Stejně jako v prostředí profesionálních rešeršních služeb i pro pokládání dotazů v internetových vyhledávačích je možné zhruba definovat strategii „rostoucí perly“ (citation pearl growing) a strategii „osekávání“ (successive fractions).

Při použití rešeršní strategie „rostoucí perly“ je zpočátku koeficient úplnosti nízký. Je nutné pomocí velmi konkrétních a specifických výrazů najít malý počet co nejvíce relevantních dokumentů (perel). Tuto metodu doporučuje pro vyhledávání v Googlu například Hlavenka [47, s. 32], neboť při jejím použití není nezbytně nutné procházet a analyzovat velké množství nerelevantních výsledků.

Teprve při případném zvyšování úplnosti je možné dotaz rozšiřovat – zobecňováním pojmů, přidáváním dalších synonymních výrazů (v disjunkci). Míra úplnosti

vyhledávání se zvýší, současně s přibývajícími irelevantními dokumenty však klesá přesnost.

Při použití strategie „osekávání“ je nejprve koeficient úplnosti relativně vysoký, ale vyhledaných záznamů je velmi mnoho – postupným přidáváním limitujících podmínek (např. operátor `site:`) a dalších výrazů (v konjunkci a negaci) dochází k poklesu počtu nelezených záznamů a růstu přesnosti. Tuto redukční strategii upřednostňuje Long [48, s. 36].

Google poskytl uživatelům širší možnosti při vytváření komplexnějších dotazů tím, že v lednu roku 2005 posunul limit pro maximální počet výrazů v dotazu z deseti na třicet dva [49].

Pro prostředí Googlu (a internetových vyhledávačů obecně) jsou specifické přinejmenším ještě dvě další metody, jak zvyšovat koeficient úplnosti, resp. přesnosti:

- Při vyhledávání úzce specifických výrazů může být nalezeno velmi malé nebo i žádné množství relevantních dokumentů. Je možné, že hledaná informace není zahrnuta v indexu – existuje sice na volně dostupném webu, ale v oblasti pro Google neviditelné (může se kupříkladu nacházet příliš hluboko ve struktuře webového sídla, viz kapitola 2.1). V tomto případě je možným řešením paradoxně položení spíše obecného dotazu. Ten bude mít za cíl určit jen potenciální informační zdroj (web), v něm je pak možné dotyčný dokument dohledat jiným způsobem (procházením, použitím vyhledávání dotyčného zdroje).

- Přesnost je možné zvýšit tím, že hledání bude směrováno nikoliv do prostoru webových stránek, ale do vybrané sekce některého internetového katalogu. Většina provozovatelů internetových vyhledávačů k tomuto účelu používá otevřený katalog Open Directory Project[1]. Jen malá část udržuje katalog vlastní (Yahoo!, LookSmart), ten má pak ale o to větší hodnotu. Google používá mírně upravenou verzi katalogu ODP pod názvem Google Directory[2].

[1] http://www.dmoz.org
[2] http://www.google.com/dirhp

6.5 Vliv softwarového prostředí na rešeršní činnost

Nezanedbatelný vliv na internetovou rešeršní činnost má také programové vybavení. Nejen v podobě sofistikovaných vyhledávacích aplikací (např. Copernic Agent[1]), kdy se už nejedná o klasické vyhledávání prostřednictvím GUI webového vyhledávače, ale především ve formě drobných podpůrných nástrojů. Mezi ně se řadí zejména různé doplňky internetového prohlížeče:

- nástroje umožňující rychlé zadání dotazu do vybraného vyhledávače prostřednictvím komponenty rozhraní prohlížeče – např. Google Toolbar.[2]
- nástroje pro lepší orientaci v cílovém dokumentu – např. podbarvení hledaného textu, vyhledávání regulárního výrazu
- javascriptové kódy ve formě záložek (bookmarklety), umožňující podrobnější analýzu dokumentu:
  - výpis všech hypertextových odkazů na stránce
  - abecední a frekvenční řazení slov z dokumentu
  - překlad textu pomocí propojení s některým internetovým překladačem jako např. Google Translate či Babelfish[3]
  - zobrazení různých verzí stránky z repozitáře systému Internet Archive[4]
  - informace o správci webového sídla (pomocí propojení s některým doménovým jmenným serverem)
- funkce prohlížeče spouštěné určitou kombinací pohybů myši (mouse gestures) – zobrazení zdrojového kódu, zobrazení cookies, vyhledání kurzorem označeného textu v Googlu apod.

Ačkoli tyto podpůrné nástroje přímo nesouvisí s tvorbou dotazu, mohou posloužit k výraznému urychlení a větší efektivitě rešeršního procesu.

[1] http://www.copernic.com
[2] http://toolbar.google.com
[3] http://babelfish.altavista.com
[4] http://www.archive.org

7 Analýza pokročilých funkcí

7.1 Operátory pro modifikaci dotazu

7.1.1 Synonymní operátor ~ (tilda)

Připojením znaku ~ k některým výrazům v dotazu dojde automaticky k vyhledání nejen jich samotných, ale také dalších slov s příbuzným významem. Operátor ~ tedy funguje jako alternativa ke klasickému způsobu zahrnování synonym do dotazu – disjunktivního spojování operátorem `OR`.

Přítomnost tohoto operátoru implikuje existenci určité formy řízeného slovníku, který jinak v Googlu chybí. Tento slovník je opět skrytý – nelze určit, která slova se ve slovníku vyskytují, ani jaká synonyma jsou jim přiřazena. Strukturu slovníku je možné odhadovat jen na základně experimentování a následného procházení výsledků. V porovnání s klasickými tezaury se však jedná o slovník velmi základní, pravděpodobně nehierarchický. Tabulka 21 dokumentuje některé příklady.

Dotaz	Vyhledávané synonymní výrazy
`~finance`	žádné
`~financial`	accounting, bank, capital, investment, money, mortgage, stock
`~medication`	drug, pharmacy, prescription, treatment
`~disease`	cancer, health
`~manual`	help, documentation, tutorial
`~research`	resource, statistics
`~sport`	baseball, basketball, fitness, football, news
`~travel`	Flights, holiday, hotel, tourism, vacation

Tab. 21: Funkčnost synonymního vyhledávání

Ač se jedná o slovník řízený, je řízen opět především pomocí strojů a strojové analýzy uživatelských dotazů – Google má tento způsob vytváření synonymních spojení od roku 2005 patentován [50].

Z pohledu informačního pracovníka nepředstavuje použití operátoru ~ výrazné pozitivum. Naopak, díky jeho netransparentnímu fungování by tak ztratil kontrolu nad tím, které výrazy jsou vyhledávány.

7.1.2 Operátor `site:`

Operátor `site:` je velmi dobře použitelný pro vyhledávání nejen v omezeném prostoru dané domény, ale i v konkrétním adresáři. Může tak z velké části nahradit chybějící vyhledávání v určité části daného webového sídla. Například prostřednictvím dotazu `site:mvcr.cz/rady/sdruzeni asociace` je možné vyhledat na stránkách Ministerstva vnitra ČR všechna občanská sdružení, která mají ve svém názvu termín „asociace“.

Operátor je také možné používat v negaci a disjunkci.

7.1.3 Operátor `numrange:`

Operátor umožňuje do dotazu zadat číselný interval (namísto prostého výčtu čísel), v podobě např.: `“nobel prize“ chemistry numrange:1982-1988`. Operátor existuje také ve zkrácené verzi, kdy je interval značen dvěma tečkami. Ekvivalentní dotaz by tedy byl `“nobel prize“ chemistry 1982..1988`.

Neocenitelnou roli hraje tento operátor zejména v případě jen částečně definovaných intervalů (např. `1945..`) a intervalů spojitých hodnot (`4.47..4.92`).

Interval lze do dotazu zadat též prostřednictvím textových polí v sekci Advanced Search, avšak jen v anglické verzi, v českém rozhraní toto nastavení chybí.

7.1.4 Operátory `intitle:`, `allintitle:`

Název webové stránky zpravidla úzce souvisí s jejím tématickým zaměřením a funguje tak jako zhuštěné vyjádření jejího obsahu. Google proto přikládá názvům velkou důležitost. Dokumentům, v nichž se hledané slovo objevuje v názvu (přesněji řečeno v *titulku,* elementu definovaném párovým tagem `<title></title>` ve zdrojovém kódu) automaticky přiděluje v procesu pořádání výsledků větší váhu.

Použitím operátoru `intitle:` v dotazu lze stanovit, že se daný výraz musí v titulku hledaných dokumentů povinně vyskytovat. Jeho použitím tak lze do určité míry zvýšit přesnost vyhledávání v případě hledání slov převážně obecného charakteru. Operátor `intitle:` se vztahuje pouze na jediný bezprostředně následující výraz, nikoliv na libovolný jeden ze všech následujících, jak je v některých pramenech chybně uváděno [47, s. 50].

Odvozený operátor `allintitle:` má stejnou funkci jako `intitle:`, vztahuje se ale již na všechny po něm následující výrazy. Dotaz `allintitle:alfa beta gama` tak povede k totožným výsledkům jako dotaz `intitle:alfa intitle:beta intitle:gama`. Používání operátoru `allintitle:`, není nicméně příliš doporučováno [48, s. 64]. Stejně jako v případě ostatních operátorů s předponou „all“ je jeho fungování v kombinaci s jinými operátory nestabilní a tedy problematické.

Podobně jako u běžného vyhledávání v textu, také u hledání v titulcích záleží na pořadí výrazů. Dotaz `intitle:google intitle:GUI tak` vypíše výsledky v jiném pořadí než dotaz `intitle:GUI intitle:google`.

V jedné věci se však vyhledávání v titulcích od běžného hledání v textu výrazně liší – v přístupu k diakritice. V případě operátoru `intitle:` (`allintitle:`) je totiž výraz vždy vyhledáván výhradně v tom diakritickém tvaru, v jakém byl zadán, navíc bez ohledu na použité jazykové rozhraní.

Tato striktní interpretace je tak výrazně odlišná od toho, jak Google vyhodnocuje diakritiku v jednotlivých slovech (viz kapitola 4.3) a frázích (kapitola 5.3.2). Tabulka 22 dokumentuje chování operátoru v porovnání právě s těmito případy na příkladu dotazu `lenochod tříprstý`, kdy jsou vyhledávány buď obě diakritické varianty („tříprstý“ i „triprsty“) nebo jen varianta jediná, specifikovaná v dotazu.

dotaz	anglické rozhraní	české rozhraní
	vyhledané varianty	
`lenochod triprsty`	obě	obě
`lenochod tříprstý`	obě	jediná
`lenochod "tříprstý"`	jediná	jediná
`“lenochod triprsty“`	obě	obě
`"lenochod tříprstý"`	obě	jediná
`"lenochod +tříprstý"`	jediná	jediná
`allintitle:lenochod tříprstý`	jediná	jediná
`allintitle:lenochod triprsty`	jediná	jediná

Tab. 22: Interpretace diakritiky ve výrazu zadaném prostým způsobem, zadaném jako fráze a zadaném pomocí `allintitle:`

Velký význam, který Google přikládá titulku dokumentu, má také jednu nevýhodu – lze jej zneužít. Ve sféře SEO již není titulek chápán výhradně jako symbolický nadpis, jako identifikátor obsahu stránky, ale také jako technický prostředek k jejímu maximálnímu zvýhodnění oproti ostatním stránkám, k dosažení co nejvyššího koeficientu SERP. Titulek tak může být uměle doplňován o různá klíčová slova, jimiž je hodnota stránky při hledání těchto slov zvyšována.

Tento trend je patrný především u komerčních prezentací, kde titulek často plní úlohu určitého „vývěsního štítu“ pro roboty vyhledávačů.

web	titulek
www.adaptavist.com	Home - Adaptavist.com - Confluence Solutions :: Licenses, Hosting, Plugins, Themes
www.refreshweb.com	B2B Internet Marketing Agency Search Engine Marketing Company Services Consulting Austin SEO Company: RefreshWeb
www.ucetnictvi-dane-pravo.cz	Vedení účetnictví, podvojné účetnictví i daňová evidence, daňové poradenství, účetní poradenství, účetní program Pohoda, daňové právo, mzdy, personalistika.
www.nejnakup.cz	NejNákup.cz: porovnání cen zboží, porovnání eShopů, porovnat zboží, srovnání ceny, levne, prodej, informace o nakupování na internetu
nemovitosti-reality.hyperinzerce.cz	Inzerce, bydlení, nemovitosti, reality, byty, domy, pozemky, chaty a chalupy, kancelářské, rodinné domy

Tab. 23: Vybrané příklady webů s titulky upravenými dle SEO

V těchto případech tak titulek zcela převzal roli kdysi relevantního metadatového tagu `<meta name="keywords">`, který již zcela pozbyl hodnoty a roboti vyhledávačů jej vesměs ignorují. Je otázkou, zda v budoucnu nebude mít titulek také podobný osud, pokud se Google rozhodne jeho váhu pro relevanci dokumentu snížit.

Tento přístup je také problematický vzhledem k citování elektronických informačních zdrojů. Zatímco v samotném textu, či v grafickém nadpisu v záhlaví vystupuje webové sídlo pod jedním určitým názvem, v titulku může být formálně reprezentováno názvem pozměněným, či zcela odlišným. To je například případ

zpravodajské služby z oblasti internetových vyhledávačů Pandia Search Central[1], která má v titulku pouze text „Search Engine Marketing and Internet Searching“.

7.1.5 Operátory `inurl:`, `allinurl:`

Tento operátor slouží pro vyhledávání dokumentů podle textových řetězců v jejich URL. Má ještě širší možnosti než příbuzné operátory `site:` a `filetype:`, čehož se dá v zásadě využít dvěma způsoby:

- Pro vyhledávání slov v URL webových sídel, v názvech jejich domén druhého a třetího řádu, i v jejich podadresářích. Slovy jsou zde míněny jakékoliv souvislé řetězce písmen a čísel, oddělené od sebe znaky jako například pomlčka, podtržítko, tečka apod.

- Pro vyhledávání samotných dokumentů (stránek) dle jejich souborových názvů. Jelikož tyto názvy mají zpravidla délku menší než deset znaků, je v podstatě nemožné pomocí nich vyjádřit obsah dokumentu. Tento způsob je tak vhodný pouze k hledání ustálených, často užívaných názvů – typicky „readme“, „faq“ či „about“.

Tento operátor, spolu s operátorem `intitle:` je velmi užívaný hlavně komunitou hackerů, hledajících mezery v zabezpečení webových sídel k získání citlivých informací (mailových adres, hesel apod.). Z hlediska běžného rešeršního procesu má použití `inurl:` smysl pravděpodobně jen jako záložní, pomocný nástroj pro zvýšení přesnosti vyhledávání.

Pokud však odhlédneme od čistě textových informací, může být tento operátor užitečný při vyhledávání obrazových dat prostřednictvím služby Google Images. Hledané obrázky je tak možné určit nejen názvem, ale také klíčovými slovy v jejich URL.

Stejně jako v případě titulu stránky, také jméno souboru bývá SEO specialisty záměrně prodlužováno – hlavně v případě dynamicky vytvářených webových stránek. Těm Google nepřiřazuje takovou váhu, jako souborům „fyzicky“ existujícím. Proto pomocí techniky *mod_rewrite* jsou tyto soubory navenek

[1] http://www.pandia.com

prezentovány robotům vyhledávačů (i čtenářům) jako skutečně existující samostatné HTML soubory. Jejich uměle vytvořený název bývá nejčastěji odvozen z titulku stránky, např. ve tvaru „http://www.lupa.cz/zpravicky/google-se-miliardove-zaloby-viacomu-neobava/“.

Diakritika je v případě operátoru `inurl:` interpretována stejným striktním způsobem, jako je tomu u předchozího `intitle:`.

7.1.6 Operátory `inanchor:`, `allinanchor:`

Funkce operátoru `inanchor:` je předmětem mnoha nedorozumění a chybných výkladů, pramenících z nepřesného překladu či nepochopení jeho podstaty. Operátor neslouží k vyhledávání webových stránek, v jejichž textu se nachází hypertextový odkaz se zadaným výrazem. Právě naopak – `inanchor:` vyhledá takové **stránky, na něž je odkazováno** z jiných stránek pomocí hypertextového odkazu, jehož znění, tzv. kotvu (anchor) uživatel zadal v dotazu. Mylný výklad se objevuje v mnoha zahraničních i českých zdrojích [51; 52].

Příčinou zmatečných interpretací je pravděpodobně absence jakékoliv oficiální dokumentace k tomuto operátoru na stránkách Googlu, narozdíl od operátorů `intitle:` a `inurl:`. O to pozoruhodnější je skutečnost, že Google tento pokročilý operátor nepřímo nabízí k použití ve formulářovém rozhraní sekce Advanced Search, určeném především pro nenáročné uživatele. Ve výběrovém poli Occurences je `inanchor:` skryt pod volbou „in links to the page“ (v české verzi „v odkazech na stránku“).

Částečně je příčinou omylů také užívaná terminologie. Odkazy vedoucí na určitou stránku, bývají především ve sféře SEO označovány jako její tzv. „zpětné odkazy“ (backlinks). Tento výraz však nemusí být zcela jednoznačně srozumitelný. Za významově výstižnější lze pokládat jiný termín, „příchozí odkazy“ (inbound links).

`Inanchor:` tedy vyhledává výhradně v textu příchozích odkazů. Za naprosto zásadní lze však považovat skutečnost, že Google prohledává text těchto odkazů vždy, nejen při použití operátoru `inanchor:`. Text příchozích odkazů je totiž i při běžném vyhledávání jedním z faktorů, hrajících roli při vyhodnocování dotazu (viz kapitola 2.3). Tento zdánlivě podružný fakt má v prostředí Googlu velký význam –

z pohledu vyhledávacího stroje jsou tak součástí dokumentu i texty všech odkazů na něj směřujících. Jinými slovy – kromě vlastního obsahu jsou **nedílnou součástí dokumentu také vazby, které má se svým okolím**. To je podstatná změna z pohledu tradičního chápání dokumentu.

7.1.7 Text příchozích odkazů jako specifický druh sémantických metadat

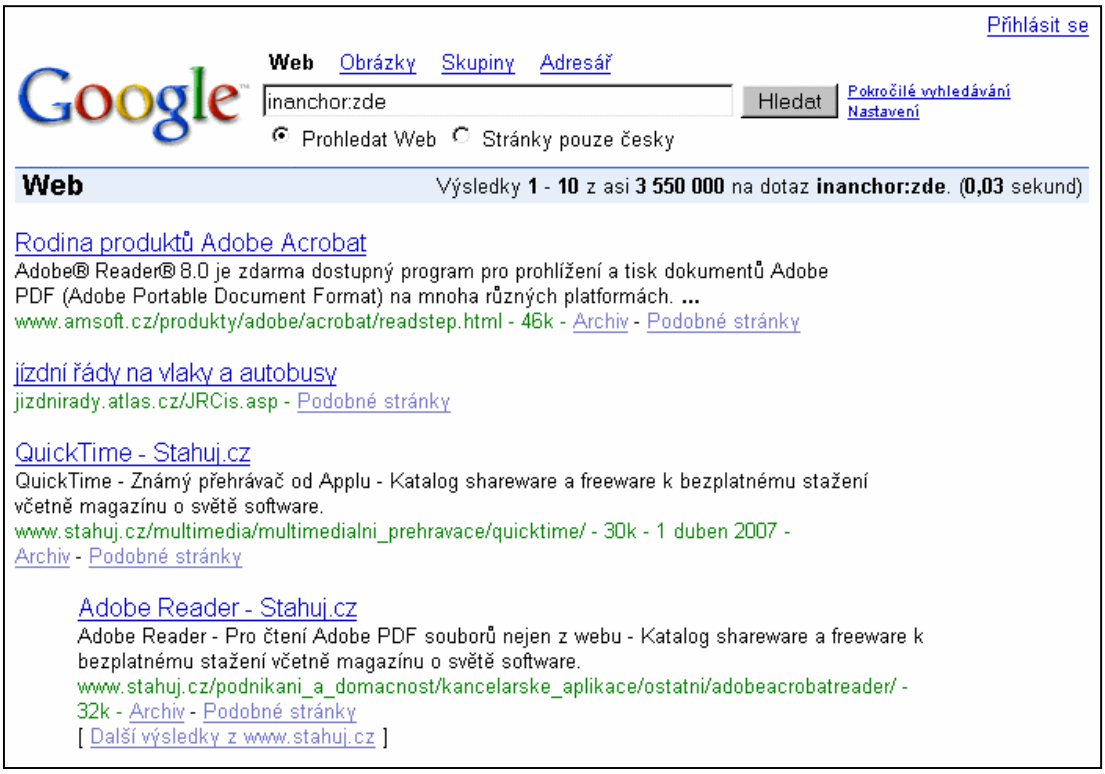
Značný význam, který Google přikládá příchozím odkazům má svůj důvod. Text, pomocí něhož se na určitý dokument odkazují (a tak ho popisují) jiné dokumenty, může poskytovat důležitou a hodnotnou informaci. Daný dokument je tak definován textovými formulacemi i jiných entit než jen samotného autora. Odkaz je tedy v podstatě doporučením, které může přinášet cenné informace o tématu a charakteristice stránky. Především pak v případech, kdy odkaz obsahuje takové výrazy, které se v odkazovaném dokumentu vůbec nevyskytují (toho lze samozřejmě i zneužít, viz dále). Například odkaz „statistika porodnosti“, který směřuje na dokument s tabulkami dat, v němž se tato slova nemusejí objevit, tak přináší novou, relevantní informaci.

Ačkoliv v teoretické rovině se jedná o zajímavou myšlenku, v praxi nicméně existuje několik důvodů, proč text v odkazech novou užitečnou informaci často nepřináší:

- Velká část hypertextových odkazů směřuje na jiné stránky jen s pomocí názvu této stránky (např. „více informací na stránkách Úřadu průmyslového vlastnictví“). To je naprosto korektní podoba odkazu, logicky však novou informaci nepřináší.

- Mnoho stránek obsahuje v odkazu formulaci podobnou této: „více informací naleznete zde“. Výraz „zde“ nepřináší již žádnou (obsahově relevantní) informaci, navíc cílové stránce přináší menší PageRank než předchozí možnost. Tato formulace je nicméně pochopitelná, neboť odpovídá přirozené konstrukci sdělení ve verbální komunikaci: „více informací naleznete zde – (následuje pokyn mluvčího)“. Také je často používána s cílem co nejvíce ulehčit orientaci v internetu méně znalým uživatelům: „chcete-li si stáhnout program IrfanView, klikněte zde“ nebo „pro odeslání objednávky pokračujte zde“.

Důkazem frekventovanosti druhého uvedeného přístupu je například výsledek dotazu `inanchor:zde`:



Obr. 8: Výsledky dotazu `inanchor:zde`

Na prvních pozicích ve výsledcích hledání se vyskytují webové stránky běžně používaných produktů a služeb, konkrétně softwaru pro čtení PDF dokumentů Adobe Reader, informačního serveru o jízdních řádech IDOS a přehrávače multimédií QuickTime. Tyto a další nalezené internetové stránky jsou klasickými zástupci velmi často navštěvovaných a také pomocí výrazu „zde“ odkazovaných webových sídel.

Pravděpodobně největší nevýhoda zahrnování příchozích odkazů do hodnocení dokumentu je přirozeně jejich „jednosměrnost“ – znění odkazu směřujícího ze stránky A na stránku B nemůže autor či správce stránky B nijak ovlivnit. Je proto jasné, že této skutečnosti lze zneužít. Pokud velké množství odkazů z různých stránek (s dostatečně velkým PageRankem) směřuje pomocí stejného výrazu na jednu cílovou stránku, je tato stránka zobrazena na čelních pozicích při vyhledávání daného výrazu, ač jej sama nikde v textu neobsahuje.

Právě na takovém principu fungují známé tzv. Google bomby. Text spojený odkazem s cílovou stránkou tak záměrně přináší mylné, či citově zabarvené informace.

7.1.8 Operátory `intext:`, `allintext:`

Operátor `intext:` je v jistém smyslu protikladem funkce `inanchor:`, neboť omezuje vyhledávání výhradně na samotné jádro dokumentu, na jeho textový obsah – nevyhledává tedy v textech příchozích odkazů, v titulcích stránek ani v jejich URL adrese. Pomocí něj je tedy možné z vyhledávání vyloučit případný nežádoucí vliv okolí (příchozí odkazy) nebo nadměrný vliv SEO (v případě marketingově upravených titulků stránek).

Stejně jako u operátorů `intitle:` a `inurl:`, i v tomto případě je česká diakritika použitá v dotazu interpretována stejným striktním způsobem.

7.2 Operátory jako samostatné funkce

7.2.1 Vyhledávání definic, operátor `define:`

Z pohledu rešeršní činnosti je velmi užitečnou funkcí Googlu vyhledávání definic nejrůznějších pojmů. To může být vhodné například při předrešeršní přípravě – pro zjištění slovní zásoby k tématu rešerše či pro nalezení dalších termínů, souvisejících s určitým pojmem.

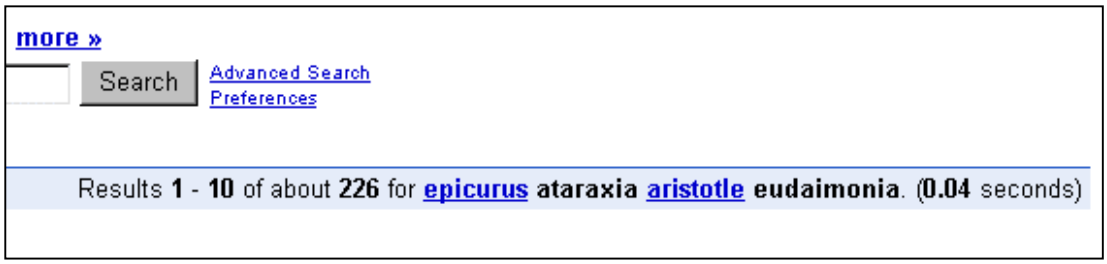
Vyhledávání definic je nicméně účinný nástroj také pro běžného uživatele, který může po Googlu požadovat především stručné a jasné odpovědi na některé dotazy, například ohledně cizích slov nebo význačných osobností. Takový uživatel pak využívá Google v podstatě jako jistý druh encyklopedie a požaduje odpovědi spíše slovníkového typu.

Definice je možné v Googlu vyhledávat v zásadě třemi způsoby:

1. Pomocí odkazů na server Answers.com, zabudovaných do výpisu výsledků
2. Pomocí operátoru `define:`
3. Pomocí zvláštních klíčových slov v dotazu

Zabudované odkazy představují nejrychlejší a nejjednodušší způsob pro zobrazení požadované definice. V záhlaví každého výpisu výsledků Googlu se nachází fráze, rekapitulující znění dotazu spolu s informací, jak velké množství výsledků bylo nalezeno. Pokud je dotaz pokládán v anglickém jazykovém rozhraní, fungují výrazy

v této frázi jako hypertextové odkazy na definice, které se nachází na encyklopedickém serveru Answers.com.



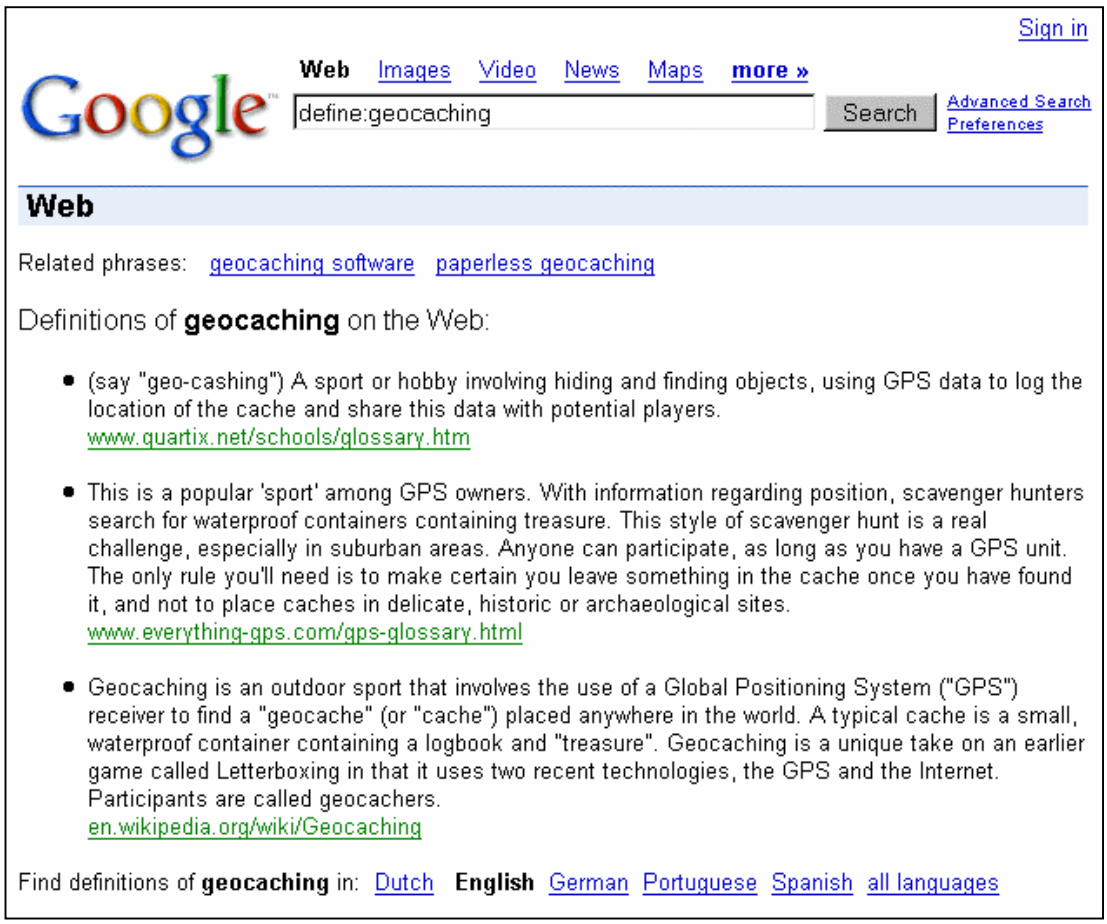
Obr. 9: Hypertextové odkazy na server Answers.com ve výpisu výsledků

Google nicméně nevytváří toto hypertextové spojení pro každý výraz, ale jen pro výrazy anglické, popřípadě pro jiné, běžně používané termíny. Proto výrazy `ataraxia` a `eudaimonia`, uvedené na obrázku 9, nefungují jako odkaz, ačkoliv server Answers.com jejich definice obsahuje.

Na rozdíl od předchozího způsobu funguje operátor `define:` ve všech jazykových rozhraních. Při jeho použití není zobrazen běžný výpis výsledků, nýbrž samostatný výpis definic z vybraných serverů. Jakým způsobem Google tyto servery pro daný dotaz vybírá, není zcela jasné. V algoritmu je pravděpodobně zohledňována především důvěryhodnost a zaměření serveru, není tedy vyloučeno, že v tomto případě do procesu vstupuje i lidský hodnotící faktor.

Kromě samotných definic, jsou uvedeny také samozřejmě jejich zdroje, dále pak slovní spojení, v nichž se výraz často na webu vyskytuje a nabídka zobrazení definic v jiném jazyce, pokud takové byly nalezeny.

Obrázek 10 ilustruje použití operátoru `define:` na příkladě výrazu `Geocaching`:



Obr. 10: Použití operátoru `define:` *pro dotaz* `geocaching`

Posledním způsobem hledání definic je vyhledávání pomocí běžného dotazu, bez použití operátorů, jde tedy o klasické vyhledávání ve výsledcích z webu. Požadavek definice určitého pojmu je v tomto případě vyjádřen specifickou formulací dotazu, např. ve tvaru `what is geocaching`. Toto vyhledávání pomocí konstruktů přirozeného jazyka bývá typické pro běžného uživatele, který nezná či nepoužívá syntaxi operátoru `define:`.

Google těmto uživatelům vychází vstříc. Pokud totiž ve znění dotazu zaznamená určitá klíčová slova (např. `define`, `definition`, `what is`), zařadí na první místo ve výpisu výsledků odkaz na definice, které by uživatel získal, pokud by použil právě operátor `define:`. Současně je zobrazen plný text první z těchto definic, jak je patrné na obrázku 11. Toto zařazování odkazů na definice je opět funkční vždy, bez ohledu na použitou jazykovou verzi rozhraní.

Obr. 11: Odkaz na definice pojmu při hledání pomocí přirozeného jazyka

V poslední době nebývale narůstá nejen v oblasti encyklopedických informací význam informačního serveru Wikipedia (česky Wikipedie). Dokumenty z této otevřené, komunitně spravované encyklopedie jsou v prostředí webových stránek často citovány a odkazovány. Proto se také velmi často objevují na předních pozicích ve výsledcích vyhledávání, zejména v případě pokládání obecných dotazů.

Některé internetové vyhledávače vyčlenily pro vyhledávání ve Wikipedii celou vlastní sekci, například český Seznam. Ve svém projektu Seznam Encyklopedie je vedle českých dokumentů z Wikipedie prohledávána také Ottova encyklopedie a internetová encyklopedie CoJeCo [53].

Google se zvláštním ošetřením wikipedických výsledků zatím jen experimentuje. V současnosti jsou v některých případech výsledky z Wikipedie předřazeny ostatním výsledkům – podobným způsobem, jako je tomu u definic na obrázku 11. Děje se tak však jen v určitých situacích, v případech použití specifických klíčových formulací. Pokusným testováním bylo zjištěno, že mezi tyto formulace patří například výrazy `define`, `what was`, `who is`, `who was`.

Paradoxně tak pouhá změna slovosledu (v případě `define`) či změna slovesného času (`what was`) rozhoduje o tom, zda bude zobrazena nabídka definic Googlu či odkaz na dokument ve Wikipedii. Obrázek 12 dokumentuje tuto změnu na příkladu stejného dotazu, jaký byl použit u obrázku 11.

Obr. 12: Odkaz na definici pojmu v encyklopedii Wikipedia

Toto předřazení výsledků z Wikipedie funguje však pouze v anglické verzi uživatelského rozhraní Googlu, ve verzi české jsou zobrazeny pouze standardní výsledky z webu.

Celkově tak vzniká poněkud chaotická situace, kterou v přehledné podobě shrnuje následující tabulka na příkladu vyhledávání definice slova `Echelon`.

Dotaz	Předřazený výsledek	Rozhraní
`definition Echelon`	získaný pomocí `define:`	libovolné
`Echelon definition`	získaný pomocí `define:`	libovolné
`define Echelon`	získaný pomocí `define:`	libovolné
`Echelon define`	z Wikipedie	jen anglické
`What is (are) Echelon`	získaný pomocí `define:`	libovolné
`What was (were) Echelon`	z Wikipedie	jen anglické
`Who is (are) Echelon`	z Wikipedie	jen anglické
`Who was (were) Echelon`	z Wikipedie	jen anglické

Tab. 24: Druh předřazeného výsledku dle výběru klíčových slov

Pro zajištění větší vypovídající hodnoty bylo kromě uvedeného výrazu testováno také chování Googlu při vyhledávání definic dalších pojmů (s odlišným jazykovým pozadím) – `ataraxia`, `jamais vu` a `red herring`, vždy se zcela totožnými výsledky.

7.2.2 Analýza okolí dokumentu, operátor `link:`

Všechny velké internetové vyhledávače disponují nějakým druhem operátoru pro zjišťování vazeb mezi indexovanými dokumenty. Google používá operátor `link:`, který slouží k vyhledávání webových stránek, obsahujících hypertextový odkaz na daný dokument (konkrétně na jeho adresu URL).

Z hlediska rešeršní činnosti tedy operátor `link:` neslouží k bezprostřednímu získávání hledaných informací. Nicméně zjišťování provázanosti cílového dokumentu s dokumenty ostatními může být přínosné ve dvou směrech:

1. Webové stránky, odkazující na určitý dokument, mohou s velkou pravděpodobností obsahovat další informace, relevantní k obsahu tohoto dokumentu.

2. Každý hypertextový odkaz na jiný dokument je ve své podstatě nejen citací, ale i doporučením. Z tohoto důvodu je možné analýzou odkazujících webových stránek (jejich počtu, obsahu a kontextu, v jakém je dokument odkazován) získat dodatečné informace o zkoumaném dokumentu – např. o jeho reputaci, důvěryhodnosti.

V porovnání s konkurenčními vyhledávači je však googlovský `link:` nástrojem s velmi omezenými možnostmi a nejistými výsledky. Příčin je několik:

- Operátor nelze v dotazu žádným způsobem kombinovat s jinými operátory. Toto omezení tak znemožňuje pokládání některých užitečných dotazů, jako je vyhledávání příchozích odkazů jen z určité domény (kombinace s operátorem `site:`) nebo vyhledávání příchozích odkazů s určitým textem (kombinace s `inanchor:`)

- Operátor vyhledává odkazy jen na jednu konkrétní zadanou adresu URL, nikoli na jakékoliv dokumenty, nacházející se v doméně. Dotaz `link:www.cas.cz` tak nalezne pouze stránky, odkazující jen na úvodní stranu webu Akademie věd ČR, nikoli však stránky odkazující např. na web Ústavu teorie informace a automatizace (www.utia.cas.cz).

- Největší nevýhodou operátoru `link:` je však neúplnost jeho výsledků, Google z neznámého důvodu uvádí ve výsledcích jen zlomek stránek, které na danou adresu URL ve skutečnosti odkazují. Po určitou dobu byl přijímán názor, že Google do výsledků nezařazuje stránky s nízkou hodnotou PageRanku a uvádí tedy jen ty významnější, postupem času však bylo prokázáno, že toto není jediným kritériem. Google nikdy oficiálně bližší podrobnosti k fungování operátoru `link:` neuvedl, pouze se vyjádřil v tom smyslu, že tento operátor vrací pouze vzorek (sampling) odkazujících stránek [54]. Sullivan se domnívá [55], že Google neuveřejňuje všechny skutečné výsledky z obavy před jejich zneužitím pomocí metod SEO.

V únoru tohoto roku nicméně Google zpřístupnil podrobnější informace o příchozích odkazech alespoň správcům webových sídel. Pokud webmaster prokáže, že je správcem určitého sídla, má pomocí nástroje Webmaster Tools[1] přístup k podrobnějším výpisům odkazů na toto sídlo. Pro běžného uživatele či rešeršéra však operátor `link:` stále zůstává nástrojem značně nedokonalým, konkurenční vyhledávače nabízejí pro prohledávání okolí dokumentů výrazně širší možnosti.

7.2.3 Analýza okolí dokumentu – srovnání s konkurencí

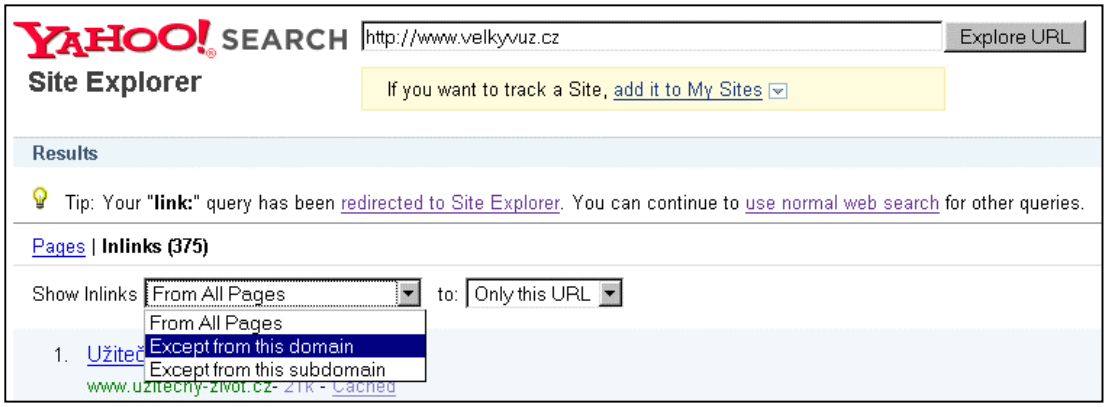
Yahoo! i Live Search disponují také operátorem `link:`, ovšem narozdíl od Googlu umožňují jeho kombinaci s dalšími operátory. Především jsou však výpisy odkazujících stránek mnohem úplnější – např. dotaz `link:www.velkyvuz.cz` vrátí v Googlu 19 výsledků, v Yahoo! celých 375.

Oba konkurenční vyhledávače provozují shodně také operátor `linkdomain:`. Ten funguje jako rozšíření předchozího operátoru – vyhledá stránky odkazující nejen na zadané URL, ale na jakékoliv URL v zadané doméně.

Yahoo! má ve svém portfoliu vyhledávacích služeb mimo jiné také nástroj Site Explorer[2], sloužící k získávání informací o indexovaných webových sídlech. Od srpna 2006 jsou všechny dotazy týkající se příchozích odkazů (`link:`, `linkdomain:`) přesměrovány právě do rozhraní tohoto nástroje [56]. Jak dokumentuje obrázek 13, prvky dialogového rozhraní umožňují nejen srozumitelné

[1] https://www.google.com/webmasters/sitemaps/
[2] http://siteexplorer.search.yahoo.com/

rozlišení operátorů `link:` („Show inlinks to: Only this URL“) a `linkdomain:` („Show inlinks to: Entire site“), ale je také možné z výsledků odfiltrovat interní odkazy, odkazující jen uvnitř zadané domény.



Obr. 13: Vyhledávání příchozích odkazů v rozhraní Yahoo! Site Explorer

Live Search disponuje velmi specifickým operátorem `linkfromdomain:`. Ten funguje jako protějšek k operátoru `linkdomain:` – nevyhledává webové stránky, odkazující na danou doménu, nýbrž naopak vypíše seznam stránek (dokumentů), na které ze zadané domény míří hypertextové odkazy. V kombinaci s dalšími operátory tak lze pokládat komplexnější dotazy – např. „jaké stránky v doméně X jsou ze stránek Y odkazovány“ (`site:X linkfromdomain:Y`).

7.2.4 Analýza okolí dokumentu – nástroje SEO

V oblasti analýzy okolí dokumentu se do značné míry protíná obor rešeršní činnosti s oborem optimalizace webových stránek pro vyhledávače. Z tohoto důvodu je možné pro analýzu okolí použít řady nástrojů, které byly primárně vyvinuty pro potřeby oboru SEO.

Typickými představiteli těchto nástrojů jsou pomůcky pro zjišťování hodnoty Google PageRanku. Pro specialistu v SEO je PageRank určité webové stránky důležitý z hlediska její potenciální návštěvnosti, pro informačního rešeršéra je zase vodítkem k míře její citovanosti. Google umožňuje uživatelům získat informaci o PageRanku určité stránky prostřednictvím Google Toolbaru, grafického doplňku rozhraní internetového prohlížeče (viz kapitola 6.5). Google Toolbar neposkytuje přímou informaci o skutečném PageRanku stránky, ale pouze takzvaný Google Toolbar PageRank (GTRP). GTPR nabývá číselných hodnot z intervalu 1 až 10 a je

zjednodušeným vyjádřením skutečného PageRanku. Obdobným způsobem funguje zjišťování míry citovanosti i u jiných vyhledávačů – např. u českého Seznamu lze jeho S-rank (hodnota 1 až 100) zjistit pomocí grafické komponenty zvané Seznam Lištička[1].

Nejsilnější nástroje pro analýzu okolí však logicky nenabízejí provozovatelé vyhledávačů, ale subjekty působící v oblasti SEO. Tou nejzákladnější pomůckou je zobrazování „ranků“ různých vyhledávačů v jediném webovém rozhraní, bez nutnosti instalovat do prohlížeče zmíněné grafické komponenty. Tuto možnost nabízí například soubor nástrojů, volně dostupných na webu i-Asap, provozovaném společností ASAP Consulting [57].

V tomto souboru se nachází také další nástroj, užitečný pro analýzu okolí dokumentu, LinkReport. Ten funguje v podstatě jako metavyhledávač pro operátor `link:` – po zadání jména domény vypíše LinkReport počet všech příchozích odkazů, které na danou doménu směřují podle vyhledávačů Google, Yahoo!, Live Search, Altavista, Morfeo a Jyxo.

Obr. 14: Výsledek hledání nástroje LinkReport pro doménu www.ikaros.cz

[1] http://www.listicka.cz

Kromě příchozích odkazů je také zkontrolována přítomnost domény v katalogu ODP. Samotný nástroj neprovádí žádné vyhledávání, pouze vypisuje počet výsledků dílčích dotazů. Ty je možné samostatně položit kliknutím na názvy vyhledávačů, které fungují jako hypertextové odkazy. Pro rešeršéra tak může LinkReport sloužit především jako „rozcestník“ do jednotlivých vyhledávačů.

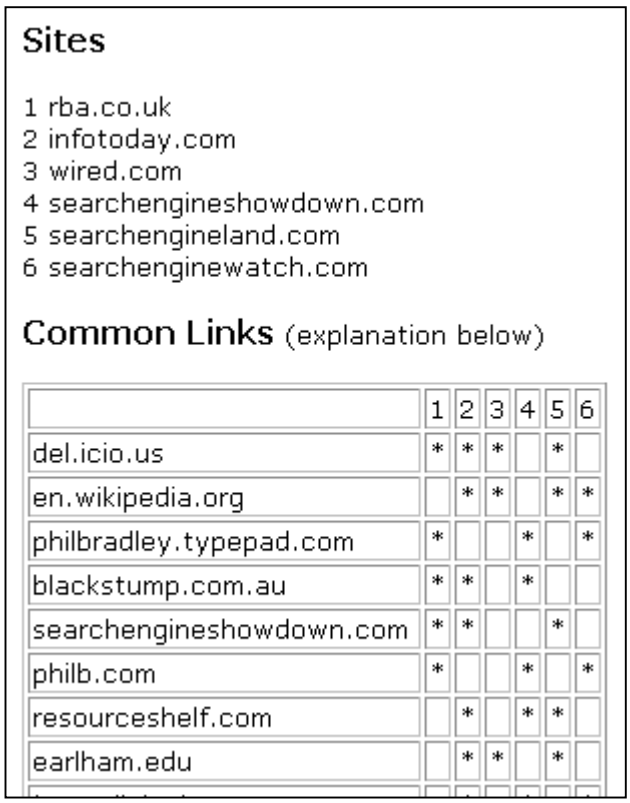
Mezi analyzovanými vyhledávači chybí Seznam, neboť jako jeden z mála žádnou formu operátoru `link:` nenabízí. Z podobného důvodu je u Live Search uveden nulový počet vyhledaných dokumentů, neboť v současné době je fungování tohoto operátoru pozastaveno (viz kapitola 5.2.3).

Mírně odlišným způsobem než LinkReport funguje nástroj Link Analyzer [58] ze serveru Bruceclay.com. Vyhledává příchozí odkazy sice jen ve třech vyhledávacích strojích (Google, Yahoo!, Live Search), ale vypisuje již přímo výsledky (v neplacené verzi nástroje jen prvních třicet, v placené dvě stě.)

Velmi zajímavý pohled na okolí stránky nabízí nástroj Common Backlinks Tool [59] ze serveru Webuildpages.com. Po zadání URL několika webových sídel jsou vyhledávány domény, z nichž vedou hypertextové odkazy právě na tyto sídla. Zjednodušeně řečeno tak lze nalézt soubor webů, které zadané URL citují a pravděpodobně se tak věnují stejné tematice jako uživatelem specifikované weby.

Určitou nevýhodou tohoto nástroje je neexistující dokumentace – není například jasné, služby kterého vyhledávače nebo vyhledávačů nástroj využívá. Mezi výhody patří například možnost zadání až deseti URL. Největším pozitivem je však aplikace jisté míry fuzzy logiky, neboť aplikace nevyhledává výhradně jen ta sídla, odkazující na všechny zadané adresy URL, ale do grafického výstupu vyhledávání jsou zahrnuty také weby, odkazující pouze na část těchto adres.

Na obrázku 15 je zobrazeno pokusné hledání webů, odkazujících na webová sídla producentů odborných časopisů z oboru online vyhledávacích služeb a také na zpravodajské weby z této oblasti.



Sites

1 rba.co.uk
2 infotoday.com
3 wired.com
4 searchengineshowdown.com
5 searchengineland.com
6 searchenginewatch.com

Common Links (explanation below)

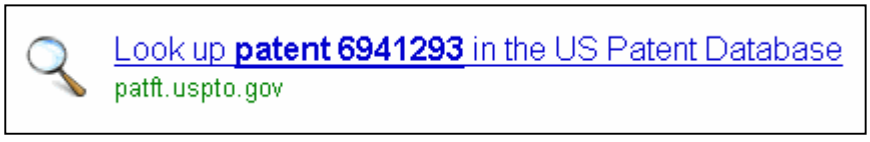
	1	2	3	4	5	6
del.icio.us	*	*	*		*	
en.wikipedia.org		*	*		*	*
philbradley.typepad.com	*			*		*
blackstump.com.au	*	*		*		
searchengineshowdown.com	*	*			*	
philb.com	*			*		*
resourceshelf.com		*		*	*	
earlham.edu		*	*		*	

Obr. 15: Výsledek hledání nástroje Common Backlinks Tool

Ve výsledcích se v popředí nacházejí obecná, tematicky nevyhraněná sídla, která ze své podstaty obsahují extrémní množství odkazů (del.icio.us, en.wikipedia.org). Na dalších pozicích se však již vyskytují weby, obsahově spřízněné s dotazovanými URL, např. weblog Phila Bradleyho či webové sídlo ResourceShelf. Celkově bylo nalezeno 88 webů, které v průměru odkazují na dvě a až tři zadané URL adresy.

7.2.5 Operátor `patent`

Kromě specializované služby Patent Search (viz kapitola 9.2.3) Google zavedl také operátor `patent`, pomocí kterého je možné provádět patenty vyhledávat také z běžného rozhraní. Pokud v dotazu figuruje tento operátor společně s číslem, které patentu přidělil americký patentový úřad (United States Patent and Trademark Office), je běžným výsledkům vyhledávání předřazen příslušný odkaz.



Obr. 16: Předřazený odkaz do databáze patentů USPTO

Operátor `patent` je funkční pouze v anglickém jazykovém rozhraní Googlu.

7.2.6 Ostatní operátory

Společným znakem dalších pokročilých operátorů, sloužících jako samostatné funkce, je jejich výrazná orientace na severoamerický prostor. Český uživatel tak dá pravděpodobně přednost tuzemským informačním zdrojům o místním počasí, programu kin apod.

Operátor `weather` zobrazí před ostatními výsledky stručnou předpověď počasí, kterou vytváří společnost Weather Underground[1]. Operátor je funkční při zadání místa na území Spojených států a dalších vybraných světových metropolí (Praha je mezi nimi). Operátor je funkční pouze v anglickém rozhraní Googlu. Jeho existence navíc v současnosti trochu pozbývá smyslu, neboť výsledky ze serveru Wunderground.com jsou i bez předřazování mezi prvními.

Informace o kurzu akcií obchodovaných na severoamerických burzách (NASDAQ, NYSE, AMEX) je možné získat použitím operátoru `stocks:`. Ten je nutné zadat v kombinaci s oficiálním symbolem (tzv. tickerem) dané akcie. Před běžnými výsledky je pak zobrazen krátký souhrn relevantních burzovních informací z beta verze služby Google Finance[2]. Operátor je opět funkční pouze v anglickém rozhraní.

Operátor `movie:` má dva způsoby použití. Pokud za ním v dotazu následuje název místa, nacházejícího se ve Spojených státech, Google zobrazí program kin nacházejících se v nejbližším okolí. Ve všech ostatních případech (resp. pokud Google výraz neidentifikuje jako místní název) jsou prohledávány recenze filmových děl, přítomné v googlovské databázi Movie Reviews[3]. Tímto způsobem je možné vyhledávat filmy například podle účinkujících herců, postav či popisu děje. Recenze je možné řadit podle relevance k dotazu, podle data či podle hodnocení filmu. Operátor `movie:` je jako jeden z mála funkční také v českém jazykovém rozhraní a také v rozhraní serveru Google.cz.

[1] http://www.wunderground.com
[2] http://finance.google.com
[3] http://www.google.com/movies

Obr. 17: Vyhledávání v databázi filmových recenzí pomocí operátoru `movie:`

Vyhledávat v telefonním seznamu bytových stanic, nacházejících se ve Spojených státech lze pomocí operátoru `phonebook:`. Za operátorem musí následovat kromě jména osoby také identifikace místa (státu unie), jinak není vrácen žádný výsledek. Vyhledávat lze i opačným směrem, specifikací telefonního čísla je možné zjistit jméno majitele stanice.

Tento nezdokumentovaný operátor původně umožňoval vyhledávání také v telefonním seznamu firem a institucí. Google však firemní seznam v tichosti bez vysvětlení stáhl a také zjišťování čísel bytových stanic postupně začleňuje do bezoperátorového vyhledávání.

Tento operátor trochu překvapivě funguje i v českém jazykovém rozhraní, včetně rozhraní serveru Google.cz.

7.3 Bezoperátorové funkce

Google nabízí uživatelům pro jejich pohodlí stále více funkcí bez toho, aby je vůbec nutil používat nějaké operátory či měnit vyhledávací služby. Uživatelova zručnost je tak nahrazována určitou formou základní umělé inteligence. Různé dodatečné informace jsou uživateli nabízeny v případech, kdy to systém uzná za vhodné, tedy pokud v dotazu automaticky rozpozná určitou konkrétní informační potřebu.

7.3.1 Začleňování výsledků ze služby Image Search

Typickým příkladem je hledání informací v grafické podobě. Pokud tělo dotazu obsahuje také určitá klíčová slova jako `image`, `picture` nebo `photo`, jsou před výsledky běžného textového vyhledávání vsunuty tři obrázky, získané službou Image Search (Hledání obrázků).

Google takto do textových výsledků zařazuje ovšem obrázky i v některých případech, kdy se v dotazu zmíněná klíčová slova nevyskytují. Lze dovodit, že tak činí podle určitého klíče u těch dotazů, u kterých pokládá grafické vyjádření jejich předmětu za přínosné, např. u dotazu `mona lisa`.

Začleňování grafických výsledků je závislé nejen na jazyku dotazu (`kůň` a `horse` zařazuje rozdílné obrázky), ale také na použitém jazykovém rozhraní. Výsledky dotazů `vánoční kapr` či `Karlův most` při použití anglického rozhraní obrazové výsledky neobsahují. Roli hraje zřejmě dokonce i velikost písmen, neboť dotaz `Mona Lisa` na rozdíl od výše uvedeného dotazu obrázky ve výsledcích nezobrazí.

Z těchto zjištění vyplývá, že obecně je zařazování relevantních výsledků (nejen) ze služby Image Search značně nekonzistentní a nelze na něj spoléhat.

7.3.2 Začleňování výsledků z dalších služeb

Velmi podobným způsobem funguje začleňování výsledků i z dalších vyhledávácích služeb – mezi nimi např. z Google Book Search a Google News. Také jimi získané informace jsou do výsledků některých dotazů zařazovány. Pravděpodobnost zařazení také zde zvyšuje použití určitých klíčových slov. Jsou jimi `book` (či `books`) a `news`.

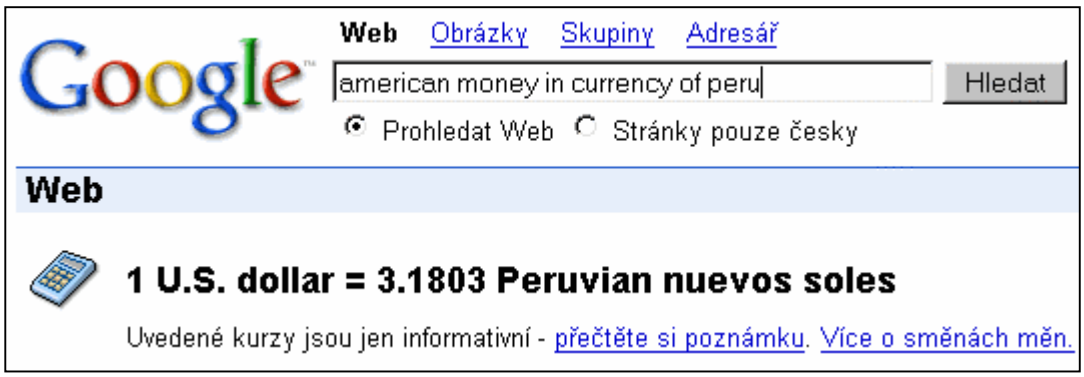
Podstatným rozdílem mezi službami z pohledu českého uživatele je různá míra jejich implementace do lokálních jazykových verzí Googlu. Zatímco výsledky z Google Book Search jsou zařazovány i v českém rozhraní (ale pouze na serveru Google.com), zpravodajské informace se zobrazují pouze v původním anglickém rozhraní. Více informací o těchto rozdílech shrnuje tabulka 15.

7.3.3 Google Calculator

Klasickou službou, dostupnou bez nutnosti užití operátorů (vyjma případných matematických) je Google Calculator. Tento nástroj disponuje kromě funkcí klasických kalkulaček také možností vzájemného převádění jednotek různých fyzikálních měr, číselných soustav a světových měn (Currency Conversion).

Podrobný popis možností Google Calculatoru je velmi často součástí nejrůznějších „návodů ke Googlu“, které kalkulátor vyzdvihují jako pozoruhodný doplněk klasického vyhledávání. Existují však některé opomíjené funkce, zajímavé z hlediska interakce uživatele s vyhledávačem.

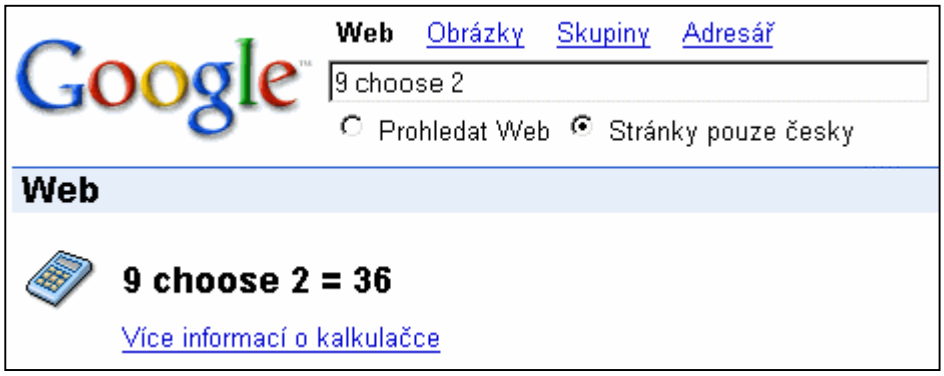
Jednou z nich je zmíněná konverze měnových jednotek. Funkce sama o sobě není zvlášť výjimečná, slouží ke zjišťování orientačních hodnot měnových kurzů, stanovených americkou bankovní společností Citibank. Možnosti specifikace obou převáděných měn jsou však velmi široké, vyhledávač totiž nevyžaduje po uživateli znalost mezinárodních měnových zkratek. Dotaz lze tak formulovat i velmi volně pomocí výrazů přirozeného jazyka – jako identifikátor měny lze použít např. výrazy `currency of Spain`, popřípadě `Spanish money`.



Obr. 18: Převod měnových jednotek

V oblasti čistě matematických funkcí je jednoduchost zadávání dotazu ještě větší přidanou hodnotou. Google Calculator tímto přívětivým způsobem umožňuje řešit úlohy z oblasti kombinatoriky [60]. Uživatel vůbec nemusí znát vzorec pro výpočet kombinací, zadá pouze celkový počet prvků v množině a počet vybíraných prvků (tedy vlastně hodnoty n a k v matematickém vzorci). Obě pak spojí klíčovým slovem `choose`.

Tato přívětivost kalkulátoru při řešení matematických úloh je o to hodnotnější, že úlohy kombinatorického typu může uživatel řešit v běžném životě velmi často. Obrázek 19 dokumentuje řešení modelové úlohy „kolik různých odstínů lze získat smícháním libovolných dvou barev z celkových devíti?“



Obr. 19: Řešení kombinatorické úlohy

8 Analýza konkurenčního prostředí

Teoreticky lze za konkurenci Googlu považovat jakýkoliv vyhledávací stroj, jehož používání uživatel zvažuje vedle používání Googlu. V praxi je však v oblasti globálně zaměřených vyhledávačů považováno za srovnatelnou konkurenci jen několik málo dalších vyhledávačů – jsou jimi především Yahoo! a Live Search, jako třetí bývá uváděn Ask.com.

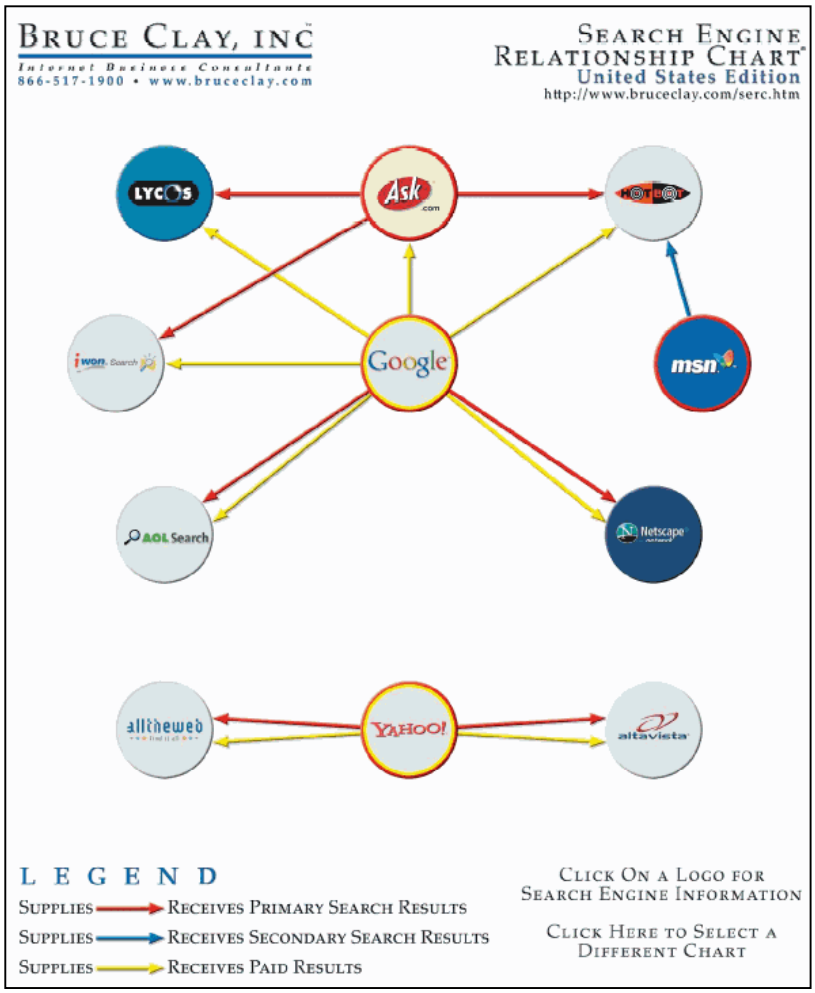
8.1 Vývoj trhu

V minulosti existoval na trhu internetových služeb velký počet komerčních subjektů, provozujících vlastní vyhledávací stroje a ještě více webových portálů, které služeb těchto strojů využívaly.

S koncem období, označovaného jako „Dot.com boom“ začalo však těchto subjektů ubývat. Zhruba mezi lety 2002–2004 postupně zcela zaniklo nebo bylo odkoupeno konkurencí velké množství internetových firem, které provozovaly takové vyhledávače, jako například Excite, AlltheWeb, Altavista, Inktomi, Teoma či WiseNut [61].

Situace v oblasti velkých hráčů na trhu se tak v posledních letech značně zklidnila, jak ilustruje i Search Engine Relationship Chart [62] od společnosti Bruce Clay, který dlouhodobě monitoruje situaci v oblasti vyhledávacích strojů.

Tento interaktivní nástroj, zachycující vzájemné vztahy mezi vyhledávači a webovými portály, ještě v polovině roku 2002 monitoroval devatenáct subjektů. V současné verzi z ledna 2007 jsou v diagramu zachyceny již jen čtyři vyhledávací stroje a sedm portálů (a vyhledávacích strojů), které výsledky vyhledávání přejímají.



Obr. 20: Diagram Search Engine Relationship Chart, verze z ledna 2007. Zdroj: Bruceclay.com

Odlišnosti na úrovni operátorů a základních funkcí Googlu a jeho konkurentů byly již popsány v příslušných předcházejících částech této práce. V následujících kapitolách budou tak konkurenční vyhledávače vymezeny vůči Googlu z hlediska obecné charakteristiky, historického vývoje a případných výrazných prvků, které ovlivňují rešeršní proces.

8.2 Yahoo!

Nejen z hlediska tržního podílu je vyhledávač Yahoo! největším konkurentem Googlu. Je nejstarším z trojice tržních lídrů a jeho nejvýraznějším specifikem je pochopitelně skutečnost, že po dlouhou dobu fungoval jako čistě katalogová služba. Hlavní devizou Yahoo! tak byl nejen odlišný způsob vyhledávání, ale také přesnější výsledky, které pramenily ze strukturovanosti prohledávaného prostoru (odkazy

zatříděné do kategorií). Postupem času začal nabízet jako alternativu ke katalogu také klasické webové vyhledávání – k němu vyžíval služeb Googlu, tehdy ještě nepřímého konkurenta.

Bylo však zřejmé, že popularita internetových vyhledávačů na úkor katalogů rychle roste. Přelomovým rokem ve vývoji Yahoo! byl proto rok 2003, kdy firma provedla několik nákladných akvizic s cílem vytvořit podmínky pro vývoj vlastního konkurenceschopného vyhledávacího stroje [63]. V dubnu tohoto roku dokončila převzetí vyhledávače Inktomi a v červenci oznámila akvizici společnosti Overture, která nedlouho předtím koupila vyhledávače AlltheWeb a Altavista.

Kombinací prvků těchto tří technologií vznikl Yahoo Search, který v únoru roku 2004 nahradil dosavadní vyhledávací služby Googlu. Vyhledávače AlltheWeb i Altavista zůstávají nadále funkční, ale jsou nyní poháněny technologií Yahoo Search. Příležitostně bývají využívány pro pokusné testování nových funkcí a služeb.

Samotný Yahoo Search[1] je jen jednou z mnoha služeb dostupných na základní stránce Yahoo![2], která se postupně proměnila v mohutný informační a zpravodajský portál. Její vizuální bohatost až přeplněnost ostře kontrastuje se střízlivým vzhledem rozhraní ostatních vyhledávačů.

Pravděpodobně nejpádnějším argumentem pro používání Yahoo Search je prostý fakt, že vrací výrazně jiné výsledky než Google. Míru překrývání výsledků u konkrétních dotazů lze zjistit pomocí metavyhledávacích nástrojů, např. Thumbshots Ranking[3].

Charakteristickým znakem, jímž se Yahoo Search odlišuje od Googlu je uplatňování tzv. *click-through popularity (CTP)* při oceňování hodnoty indexovaných dokumentů. Pojem *click-through* bývá překládán jako „proklik“, Smička [16] překládá celé spojení CTP jako „klikavost“. Jedná se o monitorování chování uživatelů poté, co jsou jim zobrazeny výsledky vyhledávání. Yahoo! zaznamenává, na které odkazy ve výsledcích uživatelé často klikají a dotyčným dokumentům zpětně zvyšuje hodnocení (Yahoo rank).

[1] http://search.yahoo.com
[2] http://www.yahoo.com
[3] http://ranking.thumbshots.com

8.3 Live Search

Microsoft vystupoval na trhu vyhledávacích služeb dlouho dobu spíše v pasivní roli. K vyhledávání pomocí svého portálu MSN Search používal služeb různých cizích technologií, konkrétně Inktomi, LookSmart a Yahoo!. Poté, co společnost Yahoo! v roce 2003 skoupila větší počet konkurenčních subjektů a začala vytvářet vlastní vyhledávač, začal také Microsoft vyvíjet svůj vyhledávací stroj. Nová verze MSN Search byla představena v listopadu 2004, v září roku 2006 byl pak vyhledávač v rámci větších koncepčních změn přejmenován na Live Search[1].

Zajímavým prvkem uživatelského rozhraní je nástroj Hodnocení výsledků (Results ranking). S jeho pomocí lze po položení dotazu dodatečně dolaďovat některé jeho parametry. Nástroj má podobu grafické lišty se třemi posuvnými jezdci.

Obr. 21: Nástroj Hodnocení výsledků

Podle toho, jak uživatel mění nastavení jednotlivých jezdců, jsou do dotazu přidávány parametry `{frsh}`, `{popl}` a `{mtch}`, společně s číselnou hodnotou z intervalu 1–100, odvislou od přesné polohy jezdce na ose. Dle nápovědy systému jsou za „oblíbené“ považovány dokumenty s velkým počtem příchozích odkazů. Definici různé míry shody (třetí jezdec) nápověda neuvádí.

Vedle funkcí a operátorů, popisovaných v jiných kapitolách této práce, lze zmínit ještě operátor `prefer:`, kterým může uživatel zvýšit váhu určitého výrazu v rámci svého dotazu.

Jako originální pokus Microsoftu o redefinici vztahu (vyhledávací) stroj-uživatel lze vnímat spuštění vyhledávače Ms. Dewey[2] na podzim roku 2006. Tato velmi okleštěná verze Live Search je k rešeršním účelům nepoužitelná, z hlediska oboru

[1] http://www.live.com
[2] http://www.msdewey.com

HCI jde nicméně o pozoruhodný projekt. Graficky velmi zdařilému uživatelskému rozhraní dominuje postava průvodkyně, slečny Deweyové (jméno nebylo zvoleno náhodou), která s uživatelem živě komunikuje a reaguje na jeho dotazy.

Obr. 22: Vyhledávací služba Ms. Dewey

8.4 Ask.com

Ačkoli je Ask.com co do velikosti tržního podílu spíše trpaslíkem, je jedním z nejstarších, dosud svébytně fungujících vyhledávačů. Mimoto obsahuje také velké množství specifických prvků, které jej odlišují od zbytku konkurence.

Ask.com si získal pozornost uživatelů brzy po svém uvedení v roce 1997, ještě pod svým původním jménem Ask Jeeves („Zeptejte se Jeevese“). Postava komorníka Jeevese, přítomná i v grafickém logu služby, měla uživatelům prezentovat uživatelsky přívětivý vyhledávač, schopný akceptovat (a věrohodně zodpovídat) dotazy položené pomocí výrazů přirozeného jazyka. Jak uvádí Sullivan [63], za touto schopností rozumět konstrukcím běžného jazyka se neskrývala žádná sofistikovaná lingvistická technologie, ale práce mnoha informačních rešeršérů. Ti pravidelně procházeli statistiky nejčastějších dotazů a vyhledávali k nim relevantní informační zdroje, které byly uživatelům ve výsledcích nabízeny.

Tento systém byl však velmi závislý na množství a kvalitě lidské práce, navíc pomíjel méně frekventované dotazy. V roce 1999 začal vyhledávač používat technologie Direct Hit, která dokumentům přisuzovala hodnotu na základě click-through popularity, četnosti prokliků. Tento systém se však neosvědčil a proto v roce 2001 Ask Jeeves koupil novou unikátní technologii vyhledávače Teoma – ten bývá

označován jako první zástupce vyhledávačů, založených na analýze shluků, tzv. *clustering search engine.*

Tato analýza je jedinečná tím, že zohledňuje obsahové zaměření indexovaných dokumentů a vytváří z nich větší tematické celky (shluky). Pro stanovení hodnoty každého dokumentu je tak důležitý nejen prostý počet příchozích odkazů, ale také to, zda odkazy přicházejí z obsahově příbuzné stránky (subject specific popularity).[1]

Vytvářené shluky jsou také hierarchicky provázány, což umožňuje vyhledávači doporučovat ke hledaným výrazům také výrazy konkrétnější, obecnější a příbuzné (levý panel na obrázku 23).

Počátkem roku 2006 došlo k přejmenování na Ask.com a v prosinci bylo představeno experimentální vyhledávací rozhraní Ask X, které přímo do výsledků začleňuje výsledky vyhledávání z různých databází (pravý panel). K dalším prvkům tohoto rozhraní patří možnost zobrazení náhledů nalezených stránek a také plynulé doporučování hledaných výrazů s tím, jak uživatel zapisuje znaky do textového pole (funkce v českých vyhledávačích známá jako „našeptávání“).

Obr. 23: Rozhraní Ask X

[1] Nelze nesrovnávat s Googlem, který svůj úspěch založil také na komplexní analýze příchozích odkazů. V případě Googlu však nehraje takovou roli tematická příbuznost odkazujícího dokumentu, jako jeho citovanost.

8.5 Ostatní formy konkurence

Konkurence v obecně definované oblasti „vyhledávání informací na internetu“ je velmi široká a různorodá. Kromě popsaných služeb (které jsou hlavními soupeři Googlu především z tržního hlediska) lze zhruba určit několik dalších skupin konkurenčních subjektů:

- Nové vyhledávače – jejich hlavní devizou není mohutnost repozitáře, ale nabízené inovace nejrůznějšího druhu. Ty mohou mít například formu specializace na jeden obor (vyhledávače weblogů, zboží, pracovních příležitostí), personalizovaného vyhledávání či prezentace výsledků ve formě grafických shluků. Za jeden z nejperspektivnějších vyhledávačů je v současnosti považován evropský Exalead, který i přes vzrůstající objem indexu stále podporuje funkce pokročilých rešeršních operátorů (viz kapitoly 5.2 a 5.3)

- Metavyhledávače – služby založené na agregaci výsledků hledání z většího počtu klasických vyhledávačů. Jejich výhodou je logicky aplikace různých vyhledávacích algoritmů na jeden dotaz, jehož výsledky jsou pak úplnější.

- Specializovaný vyhledávací software – placené produkty instalované v počítači uživatele, nabízející pokročilé metody vyhledávání a zpracování výsledků.

9 Portfolio služeb

Kromě samotného internetového vyhledávače, vlajkové lodi Googlu, tvoří dnes portfolio firmy již také velmi široká nabídka dalších služeb a produktů[1]. Řádově desítky nástrojů nabízejí služby v oblasti vyhledávání informací, jejich zpracování a pořádání a v neposlední řadě také v oblasti informační zábavy, tzv. infotainmentu.

Ač je těžištěm této práce především analýza hlavního webového vyhledávání, budou v kapitole 9.2 popsány vybrané služby, které lze považovat za informačně hodnotné a hodné pozornosti informačního specialisty.

9.1 Vývojový trend služeb internetových vyhledávačů

S tím, jak působením především generační obměny je stále větší procento populace internetově gramotné, se internet stává vskutku masovou záležitostí – a masa si žádá především zábavu.

Současný internetový infotainment však není pouze klasický pasivní „entertainment" – je charakteristický tím, že jeho hnacím motorem jsou sami uživatelé. Blogy, audioblogy (podcasts) a nyní už i videoblogy tvoří sami jejich čtenáři, posluchači a diváci, hranice mezi producenty a uživateli se smazává. Hovoří se o „médiích vytvářených uživateli", o „sociálních" či „komunitních médiích" (user generated media, social media).

Posun od chápání webového prostoru především jakožto informační pokladnice k jeho vnímání jako infotainmentového hřiště je dlouhodobě zřetelný. Google dal jasně najevo, že tento trend nepovažuje za něco podružného či podřadného. Akvizice serveru YouTube v říjnu loňského roku za částku 1,65 miliardy dolarů je výmluvným důkazem, nakolik si potenciálu komunitních médií cení. Optimisté označují akvizici jako historický milník v historii komunitních médií [64]. Ta se podle nich tímto krokem stávají plnohodnotnou konkurencí klasickým médiím a v budoucnu budou tato média vytvářená uživateli hrát ještě větší roli – v rámci vývoje webu směrem k tomu, co je označováno jako vývojový stupeň *Web 2.0*.

[1] http://www.google.com/intl/en/options/

Střízlivý pohled naproti tomu říká, že Google pouze pragmaticky zakoupil obrovský reklamní prostor. Skeptici také poukazují na fakt, že kvůli obtížně kontrolovatelnému obsahu uživatelských příspěvků na YouTube čekají Google vážné problémy v oblasti ochrany autorských práv [65].

9.2 Vybrané informační služby z nabídky Googlu

9.2.1 Google Book Search

Google Print, jak se projekt původně nazýval, byl představen v roce 2004. Jeho záměrem bylo vytvoření rozsáhlého digitálního fondu knižních děl, který měl vzniknout naskenováním knih z fondů velkých amerických (především univerzitních) knihoven. V rámci celého fondu mělo být možné plnotextové vyhledávání a po nalezení díla měl být uživateli zpřístupněn (podle stupně autorské ochrany) buď plný text, nebo případně jen jeho určitá část.

Jak konstatuje O'Leary [66], z hlediska autorských práv kontroverznímu projektu bylo za poslední roky věnováno v médiích tolik pozornosti, že samotné fungování a možnosti vyhledávání pomocí Google Book Search[1] zůstávají v odborných textech opomíjeny. O'Leary vytýká Googlu především nedostatečnou dokumentaci – kritizuje absenci informací o průběžném stavu procesu skenování v jednotlivých knihovnách, postrádá vysvětlení klíče pro výběr skenovaných titulů a seznam vydavatelů, kteří se projektu také účastní.

V důsledku dlouhodobě zvýšeného zájmu o tuto službu je nicméně dokumentace, týkající se samotného vyhledávání titulů a uživatelského rozhraní služby velmi podrobná. Jsou v ní vysvětleny důvody různého stupně zpřístupnění plných textů a zodpovězeno velké množství častých dotazů uživatelů, autorů, knihovníků i případných partnerů projektu.

Dotaz, pokládaný v Google Book Search lze směrovat buď do oblasti všech zpracovaných knihovních jednotek, pouze do oblasti plnotextově dostupných titulů nebo do databází souborných katalogů. Google totiž kromě své databáze umožňuje v rámci prohledávat i světový souborný katalog WorldCat, spolu s patnácti dalšími souborných katalogy (výčet neuvádí, je mezi nimi však i souborný katalog ČR).

[1] http://books.google.com/

Vyhledávání je možné provést také v pokročilém rozhraní, resp. pomocí operátorů. Hledaný dokument je tak možné definovat z hlediska titulu (`intitle:`), jména autora (`inauthor:`), nakladatele (`inpublisher:`), rozmezí data vydání (`date:`) a ISBN (`isbn:`). Chybí naopak možnost vyhledávat dle země vydání či jazyka dokumentu. Nicméně lze očekávat, že určitá forma geografického určení bude zavedena – do projektu je mimo amerických a britských knihoven zapojena také knihovna Madridské univerzity, Katalánská národní knihovna a nově i Bavorská státní knihovna.

Kromě standardních údajů o každé knihovní jednotce začal Google od ledna tohoto roku k vybraným titulům připojovat také informaci novou, povahově unikátní – souhrn všech geografických lokací, v dané knize zmíněných. Ty jsou navíc zobrazeny přímo v interaktivním náhledu služby Google Maps, což je jen další demonstrací synergie různých informačních služeb Googlu.

Obrázek 24 dokumentuje tuto souhrnnou geografickou informaci na příkladě knihy „Osudy dobrého vojáka Švejka“.

Obr. 24: Geografická informace ke knižnímu titulu

9.2.2 Google Scholar

Google Scholar[1], spuštěný v roce 2004 je další službou z oblasti vyhledávání dokumentů v oblasti neviditelného webu. Není však zaměřen primárně na texty knih, ale na články ve vědeckých periodikách a jiných odborných publikacích. Ty jsou často pro roboty procházející webový prostor skryty, neboť k jejich zobrazení je nutná registrace uživatele (spojená s poplatkem).

Volně dostupné zdroje zpřístupňuje Google Scholar v rozsahu plného textu, v případě placených zdrojů zobrazuje jen abstrakt. Navíc však nabízí také informaci o dostupnosti těchto dokumentů prostřednictvím služeb některé lokální odborné knihovny. Z českých institucí takto s Google Scholar spolupracují již všechny velké knihovny (NKP, STK, MZP, UJEP ad.). Sullivan uvádí [67], že tato informace o alternativní dostupnosti dokumentu může internetovou veřejnost (paradoxně prostřednictvím Googlu) opět přivést do knihoven.

Dalšími prvky, běžnými pro podobné informační služby (CiteSeet, Scirus) je citační analýza dokumentů, vyhledávání dle autora, názvu periodika, data publikace či tematického zaměření. Tematicky jsou však dokumenty roztřiďovány jen do sedmi velmi obecně stanovených kategorií:

1. Biology, Life Sciences, and Environmental Science
2. Business, Administration, Finance, and Economics
3. Chemistry and Materials Science
4. Engineering, Computer Science, and Mathematics
5. Medicine, Pharmacology, and Veterinary Science
6. Physics, Astronomy, and Planetary Science
7. Social Sciences, Arts, and Humanities

Podobně jako u Google Book Search je i této službě vytýkána absence informací o spolupracujících odborných vydavatelstvích.

[1] http://scholar.google.com/

Kromě produkce těchto vydavatelství získává Google Scholar některé články také z jiných, volně dostupných sídel. Není však jasné, podle jakého klíče takové články vybírá, které považuje za „odborné“.

Další spornou vlastností je pevný způsob řazení výstupů – Google výsledky řadí podle stejného principu, jako v případě běžného webového vyhledávání. Náročnějšímu uživateli však neumožňuje žádný jiný způsob řazení, nabízí pouze určitou formu upřednostnění novějších příspěvků (Recent articles).

9.2.3 Google Patent Search

Technické kapacity pro skenování velkého objemu tištěných titulů nejsou využívány pouze pro potřeby projektu Google Book Search. Ve spolupráci s americkým patentovým úřadem (USPTO) byla vytvořena databáze plných textů patentových dokumentů podaných ve Spojených státech v období let 1790–2006 (zhruba sedm milionů jednotek). V prosinci roku 2006 byla pak představena služba Google Patent Search[1], jejímž účelem je zpřístupnit plné texty všech patentů v rozhraní Googlu.

Patenty jsou v USA považovány (stejně jako všechny dokumenty vydané státní správou) za majetek veřejnosti (public domain). Jako takové jsou volně přístupné přímo na webových stránkách USPTO[2].

Google Patent Search má však oproti vyhledávání na webu USPTO několik výhod:

- Vyhledávání je opět optimalizované (jako ostatně u většiny služeb) pro nenáročné uživatele. V USPTO je nutné každý hledaný výraz doprovodit kódem pole, v kterém má být vyhledáván. Naproti tomu v Googlu jsou všechny výrazy v dotazu vyhledávány v kompletních plných textech všech patentů, napříč různými datovými položkami.

- V USPTO je možné v plných textech vyhledávat jen zpětně do roku 1976. Google Patent Search umožňuje plnotextové vyhledávání v celém časovém spektru až do roku 1790.

[1] http://www.google.com/patents
[2] http://www.uspto.gov/patft/

Snaha o maximální jednoduchost a nenáročnost takto prováděných patentových rešerší má nevyhnutelně i své stinné stránky:

- Služba sice disponuje rozhraním pro pokročilejší vyhledávání (a příslušnými operátory), ale možnosti rešerše v rozhraní USPTO jsou pochopitelně mnohem širší.

- Google umožňuje opět pouze jeden způsob řazení výsledků (dle svého interního mechanismu hodnocení relevance).

- Systému obsahuje (zatím) pouze databázi vydaných patentů, na webu USPTO je dostupná navíc také databáze podaných patentových přihlášek.

10 SWAT analýza Googlu

Na závěr této práce lze provést souhrnnou SWAT analýzu silných a slabých stránek Googlu, založenou také na výsledcích analýz a skutečností v této práci zjištěných.

Silné stránky

- Jednoznačně nejsilnější stránkou Googlu je v současné době jeho popularita. Ta již dosáhla takového stupně, že pro uživatele není v podstatě důležité, proč a zda se jedná o kvalitní vyhledávací stroj – Google je tak vlastně populární zejména proto, že je populární. O vysoké oblíbenosti svědčí i pronikání do běžného jazyka, sloveso „to google“ bylo v roce 2006 oficiálně zařazeno do Oxfordského slovníku.

- Ze značné oblíbenosti plyne i vysoká obchodní hodnota značky Google. Ta se podle průzkumů plynule zvyšuje, podle společnosti Interbrand [68] byla značka Google nejvíce ceněnou značkou v oboru internetových služeb za rok 2006. V celkovém pořadí se nachází na 24. pozici, což je posun o 14 příček oproti předcházejícímu roku. Především však zaznamenala nejvyšší nárůst hodnoty (o 46%) ze všech sledovaných značek.

- Ačkoli už Google neposkytuje údaj o množství indexovaných dokumentů, podle různých zdrojů se stále jedná o vyhledávač pokrývající největší počet webových dokumentů. To dále posiluje image Googlu jako „nejkvalitnějšího“ internetového vyhledávače.

- Základem úspěchu Googlu byl a stále je patentovaný PageRank, unikátní způsob stanovení informační hodnoty dokumentů na základě citační (odkazové) analýzy.

- Kromě základního webového vyhledávání disponuje celou řadou dalších služeb, specializovaných na konkrétní druhy informací (obrazové, knižní, zpravodajské, mapové, patentové, odborné aj.).

- I přes rozvoj v mnoha oblastech si Google zachovává velmi čistý a strohý ráz svého uživatelského rozhraní, které je nenáročné na hardwarové prostředky i na orientaci uživatele.

- Ve výpisu výsledků stále viditelně odděluje placené reklamní výsledky od výsledků ostatních.

Slabé stránky

- Aniž by na to Google uživatele nějakým způsobem upozorňoval, značně se liší možnosti a funkce originálního anglického rozhraní od rozhraní lokálních jazykových verzí (odlišné řazení výsledků, odlišná interpretace diakritiky, nefunkčnost některých služeb).

- Ve srovnání s ostatními vyhledávači malá nabídka pokročilých funkcí a operátorů, např. pro pravostranné rozšíření, definici váhy výrazů.

- Jen částečná aplikace operátorů booleovské logiky – nelze použít závorky pro členění dotazů.

- Netransparentně fungující stemming, vzdálenostní vyhledávání, vyhledávání synonym. U stemmingu výrazně chybí možnost jej deaktivovat.

- Nízká kvalita dokumentace u operátorů `link:`, `inanchor:` (zcela chybí), taktéž v případě stemmingu a vzdálenostního vyhledávání.

- Google často zavádí inovace v tichosti, bez oficiálního uvedení či vysvětlení. Činí tak i u změn, které mohou výrazně měnit vlastnosti a funkčnost vyhledávání (např. stemming).

- Důsledkem prohledávání extrémně velkého počtu dokumentů je jeho velmi malá přesnost. Výsledky tak často zahrnují značný počet nerelevantních dokumentů. Některé vyhledávače se snaží s tímto problémem vypořádat tak, že výsledky dále člení do tematických shluků.

- S malou přesností souvisí také určité nebezpečí pro neznalé uživatele. Velké množství výsledků i na obecné, či chybně formulované dotazy v nich může vzbuzovat pocit, že Google vždy nějaké správné výsledky najde. Jakýkoliv výsledek je totiž pro nenáročné uživatele stále především výsledkem. Uživatelé také často příliš spoléhají na to, že Google na předních pozicích zobrazuje ty objektivně nejrelevantnější výsledky – nemusejí si být vědomi vlivu SEO.

Příležitosti

- Google je finančně i technologicky natolik zajištěný, že může investovat velké částky do vývoje i v jiných oblastech – například do tvorby internetového kancelářského softwaru, mailových služeb, vývoje vlastního webového prohlížeče apod.

- Jako nejvýznamnější subjekt na trhu má Google výsadní postavení – může si dovolit určovat cenové podmínky prodeje reklamní plochy (bez ohledu na konkurenci), stanovovat trendy ve vývoji nových online služeb.

- Disponuje takovým objemem kapitálu, že nemusí nové služby za každou cenu vyvíjet – konkurenční subjekty, přicházející s určitou inovací může přímo skupovat (např. Picasa, YouTube).

Hrozby

- S tím jak narůstá velikost a rozmanitost portfolia služeb, dochází k rozmělňování jejich kvality. V určitých úzce omezených oblastech jsou specializované informační služby výkonnější.

- Riziko vzniku nové konkurence je stále vysoké – ne snad ve smyslu příchodu nového rovnocenného konkurenta, ale ve smyslu neustálého vývoje informačních služeb v dosud neprobádaných oblastech, které mohou ohrozit určitou část portfolia Googlu.

- S tím souvisí skutečnost, že Google musí vyvíjet nové a nové inovace už nikoliv primárně proto, aby nabízel uživatelům nové služby, ale proto, aby si zachoval pozici na trhu, aby konkurence nezískala náskok.

- Nejsilnější stránka Googlu, popularita, může být také jeho slabinou. V současnosti již Google v podstatě nenabízí nic, co by ostatní konkurenti (byť v různé kvalitě) nenabízeli také. Musí si proto pečlivě udržovat image novátorské, pozitivně vnímané značky. Pokud by tento image ztratil, například v důsledku vývoje některých svých kontroverzních projektů, jako je Book Search (problematika autorských práv) či Gmail (obavy ze zneužití osobních údajů), může to značně ohrozit jeho dosavadní suverénní pozici.

Závěr

Nejpodstatnějšími závěry této práce jsou samotné výsledky jednotlivých provedených analýz. Jejich společným jmenovatelem je potvrzení skutečnosti, že Google se stále více zaměřuje především na zvětšování uživatelského pohodlí. Ačkoliv to je pro uživatele jistě přínosem, daní za tento komfort je stále větší přesouvání zodpovědnosti za vyjádření informačního požadavku z uživatele na vyhledávací stroj. Ten pak po uživateli nevyžaduje tak explicitní formulaci dotazu, řídí se tacitními informacemi (odkud uživatel přichází, jaký jazyk rozhraní si nastavil, jaká hledání provedl v minulosti).

To není samozřejmě a priori negativem, naopak to může zkvalitnit vyhledávací proces a zvýšit relevanci výsledků. Na druhou stranu to však pomáhá v uživatelích pěstovat nebezpečný pocit, že výkonnost stroje je všemocná a není třeba trávit příliš času s precizní formulací dotazu – kvalitní výsledky vždy zajistí výpočetní síla a sofistikovanost algoritmů.

Z hlediska pokročilého uživatele to má za důsledek také menší kontrolu nad vyhodnocením dotazu – jazyk uživatelského rozhraní je možné ještě změnit zpět na anglický, ale například stemming vypnout nelze.

Google také pravděpodobně kvůli primární orientaci spíše na kvantitu indexovaných informací než na kvalitu vyhledávacích služeb nabízí relativně menší spektrum pokročilých vyhledávacích funkcí. Neumožňuje tak například hnízdování booleovských operací do závorek, specifikaci váhy výrazů, nespolehlivým způsobem funguje synonymní a vzdálenostní vyhledávání, operátor `link:` aj.

Řešením nedostatků Googlu, zjištěných v této práci, není nicméně ani jeho zavržení. Stále se jedná o kvalitní vyhledávač s mnoha přednostmi, který pokrývá pravděpodobně největší množinu webových dokumentů a nabízí množství dodatečných informačních služeb. Je však užitečné si být vědom jeho nedokonalostí, které tak mohou uživatele vést k hledání, nacházení a využívání také jiných rešeršních nástrojů.

Citované zdroje

1. SIMON, Herbert A. Designing Organizations for an Information-Rich World. In GREENBERGER, Martin (ed.). *Computers, Communication, and the Public Interest*. Baltimore, MD: The Johns Hopkins Press, 1971, s. 38-52.

2. KOWALSKI, Eileen. Researching with Search Engine Operators. *SEOToolSet newsletter* [online]. October 2004, vol.1 [cit. 2007-04-02]. Dostupné z WWW: <http://www.bruceclay.com/newsletter/1004/seoperators.html>.

3. NOTESS, Greg R. *Search Engine Showdown* [online]. c1999-2007, [cit. 2007-03-06]. Dostupné z WWW: <http://www.searchengineshowdown.com>.

4. BRADLEY, Phil. *Phil Bradley Home Page* [online]. [1997], last updated 2006-04-19 [cit. 2007-03-07]. Searching the Internet. Dostupné z WWW: <http://www.philb.com/searchindex.htm>.

5. LENSSEN, Philipp. *Google Blogoscoped* [online]. 2003- [cit. 2007-01-16]. Dostupné z WWW: <http://blog.outer-court.com>.

6. *Tales from the Terminal Room* [online]. Caversham : RBA Information Services, 1999- [cit. 2007-03-20]. Dostupné z WWW: <http://www.rba.co.uk/tfttr/index.shtml>. ISSN 1467-338X.

7. LIEB, Rebecca (ed.). *Search Engine Watch* [online]. 1997- [cit. 2007-01-10]. Dostupné z WWW: <http://www.searchenginewatch.com>.

8. SULLIVAN, Danny (ed.). *Search Engine Land* [online]. 2006- [cit. 2007-01-10]. Dostupné z WWW: <http://www.searchengineland.com>.

9. KOCH, Per M. (ed); KOCH Susanne (ed). *Pandia Search Engine News* [online]. c1998-2007 [cit. 2007-02-05]. Dostupné z WWW: <http://www.pandia.com/sew/index.php>.

10. PRICE, Garry (ed.). *ResourceShelf* [online]. 2001- [cit. 2007-03-07]. Dostupné z WWW: <http://www.resourceshelf.com>.

11. PAGE, Larry et al. *The PageRank Citation Ranking: Bringing Order to the Web.* Stanford, 1998. 17 s. Výzkumná zpráva. SIDL-WP-1999-0120. Department of Computer Science, Stanford University. Dostupné také z WWW: <http://dbpubs.stanford.edu:8090/pub/1999-66>.

12. BRIN, Sergey; PAGE, Larry. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the Seventh International Conference on World Wide Web, Brisbane, Australia*. 1998, s. 107-117. Dostupné také z WWW: <http://infolab.stanford.edu/~backrub/google.html>.

13. Drunk Men Work Here (Rotterdam). On Bots. In *Drunk Men Work Here* [online]. 2006-05-03 [cit. 2007-01-28]. Dostupné z WWW: <http://drunkmenworkhere.org/219>.

14. BONDAR, Serge. Search Engine Indexing Limits: Where Do the Bots Stop? In *SitePoint* [online]. 2006-04-28 [cit. 2007-01-25]. Dostupné z WWW: <http://www.sitepoint.com/article/indexing-limits-where-bots-stop>.

15. FISHKIN, Rand; POLLARD, Jeff. Google Search Engine Ranking Factors.In *SEOmoz* [online]. 2007-04-02 [cit. 2007-04-15]. Dostupné z WWW: <http://www.seomoz.org/article/search-ranking-factors>.

16. SMIČKA, Radim. *Optimalizace pro vyhledávače - SEO: jak zvýšit návštěvnost webu.* Dubany : Jaroslava Smičková, 2004. 126 s.

17. LENSSEN, Philipp. Google.com 1997-2006. In *Google Blogoscoped* [online]. 2006-04-21 [cit. 2007-01-16]. Dostupné z WWW: <http://blog.outer-court.com/archive/2006-04-21-n63.html>.

18. SULLIVAN, Danny. Google Results Different Outside The US. In *Search Engine Watch Forums* [online]. 2004-06-02 [cit. 2007-01-10]. Dostupné z WWW: <http://forums.searchenginewatch.com/showthread.php?t=60>.

19. SULLIVAN, Danny. Blair „Liar" Linkbomb Highlights Country-Specific Skewing. In *Search Engine Watch Blog* [online]. 2005-05-17 [cit. 2007-01-10]. Dostupné z WWW: <http://blog.searchenginewatch.com/blog/050517-135157>.

20. iProspect (Massachusetts). *iProspect Search Engine User Behavior Study* [online]. April 2006. [cit. 2007-01-29]. Dostupné z WWW: <http://www.iprospect.com/premiumPDFs/WhitePaper_2006_SearchEngineUserBehavior.pdf>.

21. NOTESS, Greg R. Search Engine Showdown Analysis: Boolean Searching on Google. In *Search Engine Showdown* [online]. 2000-11-03, [cit. 2007-03-06]. Dostupné z WWW: <http://www.searchengineshowdown.com/features/google/googleboolean.html>.

22. Seznam.cz (Praha). *Seznam Nápověda* [online]. c1996-2007 [cit. 2007-03-07]. Pokročilé zadávání dotazu. Dostupné z WWW: <http://napoveda.seznam.cz/cz/pokrocile-zadavani-dotazu.html>.

23. Centrum.cz (Praha). *Nápověda Centrum.cz* [online]. c1999-2007 [cit. 2007-03-07]. Pokročilé dotazy (pokročilá syntaxe). Dostupné z WWW: <http://napoveda.centrum.cz/index.php?root=156>.

24. POKORNÝ, Jaroslav et al. *Dokumentografické informační systémy.* 1. Vyd. Praha : Karolinum, 1998. 5.1 Boolovský model, s. 83-89.

25. BLAKEMAN, Karen. Google expands and turns scholar. *Tales from the Terminal Room* [online]. November 2004, no. 57 [cit. 2007-03-30]. Dostupné z WWW: <http://www.rba.co.uk/tfttr/archives/2004/nov2004.shtml>. ISSN 1467-338X.

26. BLAKEMAN, Karen. Specialist Tools for Tackling the Hidden Web. In *Sborník příspěvků konference Inforum 2006* [online]. Praha: VŠE, 2006. Dostupné z WWW: <http://inforum.cz/inforum2006/prispevek.php-prispevek=81.htm>.

27. Google (Mountain View, CA). *Google Help : Basics of Search* [online]. c2007 [cit. 2007-03-07]. Word variations (stemming). Dostupné z WWW: <http://www.google.com/help/basics.html>.

28. BLACHMAN, Nancy; PEEK, Jerry. *Google Guide* [online]. [2002] [cit. 2007-03-07]. Dostupné z WWW: <http://www.googleguide.com>.

29. NOTESS, Greg R. Google Starts Auto Stemming Searches. In *Search Engine Showdown* [online]. 2003-11-28 [cit. 2007-03-07]. Dostupné z WWW: <http://www.searchengineshowdown.com/blog/2003/11/google_starts_auto_stemming_se.shtml>.

30. STROSSA, Petr. Český lemmatizátor: Proč a hlavně jak? In *Science World* [online]. 2002-08-26 [cit. 2006-11-07]. Dostupné z WWW: <http://scienceworld.cz/sw.nsf/0/6B67159F988B95AFC1256E970048C2BB>.

31. PROKOP, Marek. Dušan Janovský: Seznam se snaží být přísný. In *Vyhledávače.info* [online]. 2006-03-13 [cit. 2007-03-07]. Dostupné z WWW: <http://vyhledavace.info/clanky/187/janovsky-o-seznamu>.

32. SEIDMAN, Eytan. We are flattered, but.... In *Live Search's WebLog* [online]. 2007-03-28 [cit. 2007-03-30]. Dostupné z WWW: <http://blogs.msdn.com/livesearch/archive/2007/03/28/we-are-flattered-but.aspx>.

33. VINE, Rita. The Experts Declare „Nothing New" on Google Proximity Searching. In *SiteLines* [online]. 2003-08-11 [cit. 2007-03-13]. Dostupné z WWW: <http://www.workingfaster.com/sitelines/archives/2003_08.html#000096>.

34. ALSHAWI, Hiyan. Fill in the blanks. In *Official Google Blog* [online]. 2005-08-09 [cit. 2007-03-13]. Dostupné z WWW: <http://googleblog.blogspot.com/2005/08/fill-in-blanks.html>.

35. Google (Mountain View, CA). *Google: Web Search Help Center* [online]. c2007 [cit. 2007-03-13]. Using Web Search. Search Tips. Does Google support wildcard searches? Dostupné z WWW: <http://www.google.com/support/bin/answer.py?answer=3178>.

36. NOTESS, Greg R. Unusual Power Web Searching Commands. *Online* [online]. 2003, vol. 27, no. 6 [cit. 2006-03-13]. Dostupné z WWW: <http://www.infotoday.com/online/nov03/OnTheNet.shtml>. ISSN 0146-5422.

37. KOSTOFF, Ronald N.; RIGSBY, John T.; BARTH, Ryan B. Adjacency and Proximity Searching in the Science Citation Index and Google. *Journal of Information Science.* 2006, vol. 32, no. 6, s. 581-587.

38. SHAY, Kevin. *Google API Proximity Search (GAPS)* [online]. Version 1.1. 2002 [cit. 2007-03-14]. Dostupné z WWW: <http://www.staggernation.com/cgi-bin/gaps.cgi>.

39. NOTESS, Greg R. Search Engine Update. *Online.* Sep/Oct 2006, vol.30, iss. 5, s 12. Dostupné také z WWW: <http://proquest.umi.com/pqdweb?did=1128110841&sid=5&Fmt=3&clientId=45145&RQT=309&Vname=PQD>. ISSN 0146-5422.

40. Exalead. *Exalead - Internet search engine, image search engine, video search engine* ... [online]. c2000-2007 [cit. 2007-03-30] Exalead Web Search Features. Dostupné z WWW: <http://www.exalead.com/search/querySyntaxReference>.

41. BLAKEMAN, Karen. Exalead enhances NEAR command. *Tales from the Terminal Room* [online]. March 2006, no. 69 [cit. 2007-03-30]. Dostupné z WWW: <http://www.rba.co.uk/tfttr/archives/2006/mar2006.shtml>. ISSN 1467-338X.

42. HARRIS, Gwen. WSG Search Engine Comparison Chart. In *Web Search Guide* [online]. 2006-10-06, page updated 2006-11-03 [cit. 2007-03-07]. Dostupné z WWW: <http://www.websearchguide.ca/research/searcomp2.htm>.

43. SKLENÁK, Vilém. Metadata, sémantika a sémantický web. In *Sborník příspěvků konference Inforum 2004* [online]. Praha: Vysoká škola ekonomická, 2004. Dostupné z WWW: <http://www.inforum.cz/inforum2004/prispevek.php-prispevek=58.htm>.

44. PAPÍK, Richard. Vyhledávání informací I. Umění či věda? *Národní knihovna*. 2001, roč.12, č. 1, s. 18-25. Dostupné také z WWW: <http://full.nkp.cz/nkkr/NKKR0101/0101018.html>.

45. KUČEROVÁ, Helena. *Metody a techniky vyhledávání informací* [online]. Praha : Vyšší odborná škola informačních služeb, [2005], datum poslední aktualizace 7. 1. 2006 [cit. 2007-03-20]. Vyhledávání informací. Dostupné z WWW: <http://web.sks.cz/users/ku/MTI/vyhled.htm>.

46. JACKSON, Peter. Artificial Intelligence and Information Retrieval. *Searcher*. Jan 2005, vol. 13, iss. 1, s. 29-33. Dostupné také z WWW: <http://proquest.umi.com/pqdweb?did=780119021&sid=8&Fmt=4&clientId=45145&RQT=309&Vname=PQD>. ISSN 1070-4795.

47. HLAVENKA, Jiří. *Mistrovství ve vyhledávání na Internetu.* 2. Aktualiz. Vyd. Brno : Computer Press, 2004. 196 s.

48. LONG, Johnny. *Google hacking.* Brno : Zoner Press, 2005. 472 s.

49. LENSSEN, Philipp. Google Raises Word Limit to 32. In *Google Blogoscoped* [online]. 2005-01-22 [cit. 2006-12-10]. Dostupné z WWW: <http://blog.outer-court.com/archive/2005-01-22-n48.html>.

50. Google (US). Methods and apparatus for determining equivalent descriptions for an information need. Jeffrey A. Dean aj. (inventors). MPT6: G06F17/30. USPTO Patent, 6941293. 2005-09-6.

51. PEČENÝ, Ondřej. Google pro pokročilé. In *Sborník příspěvků konference Inforum 2004* [online]. Praha: Vysoká škola ekonomická, 2004. Dostupné z WWW: <http://www.inforum.cz/inforum2004/prispevek.php-prispevek=63.htm>.

52. ISKRA, Jiří. *Google : vyhledávání, Gmail, Google Talk a další služby.* Brno : Computer Press, 2006. 143 s.

53. Seznam.cz (Praha). *Seznam Nápověda* [online]. c1996-2007 [cit. 2007-03-031] Hledání v Encyklopedii. Dostupné z WWW: <http://napoveda.seznam.cz/cz/hledani-v-encyklopedii.html>.

54. Google (Mountain View, CA). *Webmaster Help Center* [online]. c2007 [cit. 2007-04-02]. How do I find out what sites link to me? Dostupné z WWW: <http://www.google.com/support/webmasters/bin/answer.py?answer=34453>.

55. SULLIVAN, Danny. Google Releases New Link Reporting Tools. In *Search Engine Land* [online]. 2007-02-05 [cit. 2007-04-02]. Dostupné z WWW: <http://searchengineland.com/070205-165836.php>.

56. GARG, Priyank. Pointing Webmaster Queries to Site Explorer. In *Yahoo! Search Blog* [online]. 2006-08-10 [cit. 2007-04-03]. Dostupné z WWW: <http://www.ysearchblog.com/archives/000343.html>.

57. ASAP Consulting (Praha). *I-Asap* [online]. [2005] [cit. 2007-04-05]. Nástroje. Dostupné z WWW: <http://www.i-asap.net/nastroje.php>.

58. Bruce Clay (Moorpark, CA). *Bruceclay.com* [online]. c1996-2007 [cit. 2007-04-05]. Search Engine Optimization. Dostupné z WWW: <http://www.bruceclay.com/web_rank.htm>.

59. We Build Pages (Troy, NY). *We Build Pages* [online]. c1999-2007 [cit. 2007-04-06]. Common Backlinks Tool. Dostupné z WWW: <http://www.webuildpages.com/tools/common-backlinks>.

60. Google (Mountain View, CA). *Google: Web Search Features* [online]. c2007 [cit. 2007-04-08]. How to use the Google calculator. Dostupné z WWW: <http://www.google.com/help/calculator.html>.

61. NOTESS, Greg R. *Search Engine Showdown* [online]. 1999, last updated 2006-09-17 [cit. 2007-04-08]. Search Engine Showdown Reviews. Dostupné z WWW: <http://www.searchengineshowdown.com/reviews/>.

62. Bruce Clay (Moorpark, CA). *Bruceclay.com* [online]. c1996-2007 [cit. 2007-04-05]. Search Engine Relationship Chart. Dostupné z WWW: <http://www.bruceclay.com/searchenginerelationshipchart.htm>.

63. SULLIVAN, Danny. Major Search Engines and Directories. In *Search Engine Watch* [online]. 2004-04-28 [cit. 2007-02-10]. Dostupné z WWW: <http://searchenginewatch.com/showPage.html?page=2156221>.

64. WEIL, Debbie. We‘re at the inflection point for corporate blogging. In *BlogWrite for CEOs* [online]. 2006-10-26 [cit. 2007-01-15]. Dostupné z WWW: <http://www.blogwriteforceos.com/blogwrite/2006/10/the_inflection_.html>.

65. INGRAM, Mathew. So what’s next for YouTube? In *Mathewingram.com/work* [online]. 2006-10-26 [cit. 2007-01-15]. Dostupné z WWW: <http://www.mathewingram.com/work/2006/10/10/so-whats-next-for-youtube/>.

66. O‘LEARY, Mick. Google Book Search Has Far to Go. *Information Today*. Nov. 2006, vol. 23, no. 10 [cit. 2007-01-15]. Dostupné také z WWW: <http://www.infotoday.com/it/nov06/Oleary.shtml>. ISSN 8755-6286.

67. SULLIVAN, Danny. Google Scholar Offers Access To Academic Information. In *Search Engine Watch* [online]. 2004-11-18 [cit. 2007-04-11]. Dostupné z WWW: <http://searchenginewatch.com/showPage.html?page=3437471>.

68. Interbrand (New York). *Best Global Brands 2006* [online]. July 2006 [cit. 2007-04-15]. Dostupné z WWW: <http://www.ourfishbowl.com/images/surveys/BGB06Report_072706.pdf>.

Další použité zdroje

ANTHES, Gary. Shrinking Degrees of Separation. In *Computer World* [online]. 2005-10-24 [cit. 2007-02-07]. Dostupné z WWW: <http://www.computerworld.com/printthis/2005/0,4814,105595,00.html>.

BARKER, Joe. *The Library-University of California, Berkeley* [online]. Berkeley : University of California Berkeley Library, c2006 [cit. 2007-03-07]. Finding Information on the Internet : A Tutorial. Dostupné z WWW: <http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/FindInfo.html>.

BATELLE, John. The Birth of Google. *Wired Magazine* [online]. August 2005, iss. 13.08 [cit. 2007-04-06]. Dostupné z WWW: <http://www.wired.com/wired/archive/13.08/battelle.html?tw=wn_tophead_4>. ISSN 1059-1028.

BRADLEY, Phil. Search Engines: No Longer About Search. *Ariadne* [online]. 2005, is. 42 [cit. 2006-06-14]. Dostupné z WWW: <http://www.ariadne.ac.uk/issue42/search-engines>. ISSN 1361-3200.

DUNN, Ross. Examining Yahoo!'s Site Explorer Tool: How it Can Help You. In *Search Engine Guide* [online]. 2006-10-18 [cit. 2007-03-30]. Dostupné z WWW: <http://www.searchengineguide.com/dunn/2006/0818_rd1.html>.

EDMONDS, Andy. Search Macros: LinkfromDomain. In *Live Search‘s WebLog* [online]. 2006-10-16 [cit. 2007-03-30]. Dostupné z WWW: <http://blogs.msdn.com/livesearch/archive/2006/10/16/search-macros-linkfromdomain.aspx>.

HARRIS, Gwen. *Web Search Guide* [online]. c2002-2007 [cit. 2007-03-07]. Dostupné z WWW: <http://www.websearchguide.ca>

HOUSER, Pavel. Lemmatizace a české fulltexty. In *Science World* [online]. 2002-11-14 [cit. 2006-11-07]. Dostupné z WWW: < http://www.scienceworld.cz/sw.nsf/ID/ABB2F3872D27F01DC1256E970048C2ED>

KTD : Česká terminologická databáze knihovnictví a informační vědy (TDKIV) [online]. Praha : Národní knihovna České republiky, 2003. Dostupné z WWW: <http://sigma.nkp.cz/cze/ktd>.

NOTESS, Greg R. Google Inconsistencies. In *Search Engine Showdown* [online]. 2000, last updated 2003-11-10 [cit. 2007-03-06]. Dostupné z WWW: <http://www.searchengineshowdown.com/features/google/inconsistent.shtml>.

PAPÍK, R.; MICHALÍK, P.; MICHALÍK, P.; NOVÁČEK, L. *Internet - ekonomické, marketingové a finanční aplikace : strategie vyhledávání a prezentace.* Praha : Ekopress, 1998. 220 s.

PIPER, Paul S. Google Spawn: The Culture Surrounding Google. *Searcher* [online]. June 2004, vol. 12, no. 6, s. 26-33. Dostupné z WWW:<http://www.infotoday.com/searcher/jun04/piper.shtml>. ISSN 1070-4795.

RASHMI, Sinha. Google's pragmatic, data-driven approach to user interface design. In *Rashmi Sinha's weblog* [online]. 2005-01-13 [cit. 2007-01-16]. Dostupné z WWW: <http://www.rashmisinha.com/archives/05_01/googles-pragmatic-datadriven-approach-to-user-interface-design.html>.

SADEH, Tamar. Google Scholar Versus Metasearch Systems. *High Energy Physics Libraries Webzine* [online]. March 2006, iss. 12 [cit. 2007-04-11]. Dostupné z WWW:<http://library.cern.ch/HEPLW/12/papers/1/>. ISSN 1424-2729.

SKLENÁK, Vilém a kol. *Data, informace, znalosti a Internet.* 1. vyd. Praha : C. H. Beck, 2001. 507 s.

SONNENREICH, Wes; MACINTA, Tim. *Web Developer.com Guide to Search Engine.* Indianapolis, IN : Wiley Publishing, 1998. Chapter 1: A History of Search Engines. Dostupné také z WWW: <http://www.wiley.com/legacy/compbooks/sonnenreich/history.html>.

SPENCER, Stephan. *Unlocking Google's Hidden Potential* [online]. 2006 [cit. 2007-03-22]. Dostupné z WWW: <http://www.netconcepts.com/learn/google-ebook.pdf>.

SULLIVAN, Danny. Google Launches Google Patents, Full-Text US Patent Searching. In *Search Engine Land* [online]. 2006-12-13 [cit. 2007-04-11]. Dostupné z WWW: <http://searchengineland.com/061213-200005.php>.

WALL, Aaron. History of Search Engines: From 1945 to Google 2007. In *Search Marketing Info* [online]. c2004, last modified 2007-04-14 [cit. 2004-04-18]. Dostupné z WWW: <http://www.search-marketing.info/search-engine-history/>.

WILLIAMSON, Alan. An evening with Googles Marissa Mayer. In *Alan Williamson* [online]. 2005-01-12 [cit. 2007-01-16]. Dostupné z WWW: <http://alan.blog-city.com/an_evening_with_googles_marissa_mayer.htm>.

Evidence výpůjček

Prohlášení:

Dávám svolení k půjčování této diplomové práce. Uživatel potvrzuje svým podpisem, že bude tuto práci řádně citovat v seznamu použité literatury.

V Praze, 23. 4. 2007.

Vojtěch Sedláček ( )

Jméno	Katedra / Pracoviště	Datum	Podpis