Google pro pokročilé

Ondřej PEČENÝ
Ústav informačních studií a knihovnictví,
Filozofická fakulta, Univerzita Karlova v Praze

INFORUM 2004: 10. konference o profesionálních informačních zdrojích
Praha, 25. - 27.5. 2004

Abstrakt. Jak mohou informační specialisté „přemoci“ fenomén zvaný Google? Z velké části jeho vlastními zbraněmi. Stačí se naučit dokonale využívat všech jeho vyhledávacích schopností. Naprostá většina jeho uživatelů nepoužívá formulaci dotazu pomocí pokročilých technik. Cílem tohoto příspěvku je seznámit se způsoby vyhledávání v Googlu, které poskytují informačním specialistům výhodu oproti běžným uživatelům. Jak například používat Google pro vyhledávání v oblasti českého internetu? Jak si nechat Googlem přeložit webovou stránku z jednoho světového jazyka do druhého? Google neslouží pouze lidem s dobrými úmysly. Je možno jej použít několika způsoby jako zbraň (neoprávněné získání přístupu k databázím, prozrazení hesel, nelegální software, Google-bombing atd.). Googlemanie – nebezpečí představy, že když „to není v Googlu, není to nikde na internetu“. Google není pouze vyhledávačem webových stránek. Jaké další služby poskytuje?

Minulý rok byli informační specialisté na této konferenci Mary Ellen Bates vyzváni, aby „bojovali“ proti Googlu. V tomto příspěvku navazujícím na můj loňský a rozšiřujícící jej byste se měli blíže dozvědět, jaké techniky přitom mohou používat. Základní taktikou bude pokud možno dokonalé využití všech vyhledávacích schopností samotného serveru Google. Vždyť technik pokročilého vyhledávání zde využívají pouhá 2% běžných uživatelů, jak loni M. E. Bates citovala jednoho ze zakladatelů Googlu. Nejprve krátce k popularitě Googlu mezi fulltextovými vyhledávacími službami:

Google i v roce 2003 zvítězil jako značka s největším vlivem na životy více než 4000 respondentů serveru Brandchannel.com z více než 85 zemí. Předstihl tak např. značku Apple, Mini, Coca-Cola, Samsung, Ikea či Nokia. Je třeba samozřejmě přihlédnout k tomu, že průzkum probíhal mezi lidmi pohybujícími se na internetu.

Co k obrovské popularitě Googlu vede? Bezesporu je to především obrovská velikost databáze indexovaných i neindexovaných dokumentů na Webu. K 23.4.2004 jich bylo podle úvodní stránky 4.285.199.774. Uživatelé rovněž oceňují jednoduché uživatelsky přívětivé rozhraní, za kterým se skrývá technologicky sofistikovaný vyhledávací systém a absenci agresivních reklam. Google, narozdíl od jiných vyhledávačů, které se často vivinuly v portály, zůstává kompletně zaměřen pouze na vyhledávání.

Jak je Google používán česky mluvícími uživateli internetu? Velkou předností je pro ně skutečnost, že Google používá v závislosti na jazykovém nastavení operačního systému české rozhraní pro vyhledávání. Komu by nevyhovovalo, může si zvolit z dalších 87 jazyků, mezi kterými nechybí např. esperanto, latina, nepálština či hackerština. Pod vyhledávacím polem je možnost zaškrtnout „Vyhledávání stránek česky“. Tato funkce však nefunguje na sto procent a osobně bych doporučoval ji používat v omezené míře. Někdy se totiž při jejím použití objevují mezi vyhledanými stránkami i ty, jejichž jazyk nepreferujeme. Kromě češtiny můžete vyhledávat stránky ve 34 dalších jazycích. Stránky českého internetu je možno často s úspěchem vyhledat pomocí příkazu „site“, který se pak spolu s vyhledávaným termínem zadává ve tvaru „site:.cz“. Problém nastává, pokud nejsou hledané stránky umístěné v doméně CZ, s čímž se setkáváme i vzhledem k cenám registrace jednotlivých domén stále častěji. Pro srovnání, dvouletá registrace domény 2. úrovně v TLD (Top Level Domain = doména nejvyšší úrovně) CZ stojí cca 1600,- Kč bez DPH a v doméně COM, NET či ORG pouze 400,- Kč.

Ve srovnání s českými vyhledávači Google plně obstojí, i když naprostá většina českých uživatelů internetu používá zřejmě Seznam. Je možno to zjistit z analýzy statistik serveru Navrcholu.cz z února 2004. Seznam obsahuje lepší hierarchicky tříděný katalog stránek pro českou oblast internetu. Používal však částečně technologii Googlu. Nyní využívá Jyxo. Google využívá katalogu Open Directory Project, který přebírají i další vyhledávače, ale ten je českými uživateli teprve objevován.

Google rozlišuje mezi slovy zadávanými s diakritikou nebo bez ní a pro nalezení většího počtu stránek je někdy nutno (oproti českému vyhledávači Jyxo či Morfeo) zadávat slova skloňovaná nebo časovaná. I v angličtině je nalezen jiný počet stránek např. při zadání slova „airline“ a „airlines“. Google tedy nepoužívá lematizaci (angl. stemming).

Informační pracovníci v komerčních i nekomerčních institucích by rozhodně měli využívat všech možností, které jim nabízí pokročilé (rozšířené) vyhledávání na Googlu. Kromě běžného vyhledávání pomocí logických operátorů zde můžeme vyhledávat přesné fráze, či stránky aktualizované v zadaném (předvoleném) časovém období či omezovat vyhledávání pouze na určitý formát souboru. V současnosti je k dispozici vyhledávání souborů typu PDF, PS, DOC, XLS, PPT a RTF, ale pomocí příkazu ve tvaru „filetype:ext“, kde „ext“ je zvolenou koncovkou souboru, je možno vyhledávat i soubory typu TXT, HTM, HTML, ASP, PHP, WPD, WML, CDR atd. Vyhledané dokumenty nám Google pomocí odkazu „View HTML“ umožní alespoň v základní podobě zobrazit aniž bychom k tomu potřebovali příslušný program. Tento odkaz může sloužit i pro rychlý náhled na dokument bez nutnosti jej stahovat v originálním formátu. Vyhledávané termíny jsou v tomto případě v zobrazeném dokumentu pro lepší orientaci barevně odlišeny. Google v současnosti obsahuje odkazy na zhruba 35.000.000 dokumentů jiného formátu než HTML, což je oblast nazývaná mnohými „neviditelný web“, v angličtině „invisible“, „deep“ či „dark web“. Více se tímto tématem ve svém příspěvku zabýval např. Martin Lhoták na Inforu 2002. Sám považuji takto pojímaný termín „neviditelný web“ za dosti subjektivní záležitost odvíjející se například právě od znalosti pokročilého vyhledávání. Mám tím na mysli to, že míra „neviditelnosti“ dokumentů jiných než HTML může být u každého dána schopností je nalézt (třeba pomocí Googlu). „Neviditelný web“ se však netýká jen těchto dokumentů.

Kromě toho je možno na stránce pokročilého vyhledávání zvolit místo výskytu vyhledávaných termínů. Je možno je vyhledávat kdekoli na stránce, v jejím názvu, těle, v adrese URL nebo v odkazech. Vyhledávání můžeme omezit jen na doménu libovolného stupně (např. admin.psisco.cz, inforum.cz, cz, com nebo org). Můžeme se pokusit najít stránky podobné nějaké zadané stránce nebo ty, které na ni odkazují. Stránka pokročilého vyhledávání nabízí možnost využívání všech těchto funkcí, pokud si nechcete pamatovat konkrétní příkazy zadávané do políčka jednoduchého vyhledávání (ty mohou ale práci zrychlit). Následuje seznam použitelných příkazů s příklady:

josef +I. - „I“ musí být ve vyhledaných stránkách obsaženo (stop word)
"zákon o účetnictví" - najde přesnou frázi
"ing. * liebich" - najde přesnou frázi s různými slovy místo hvězdičky
latimeria OR latimerie - najde stránky s minimálně jedním termínem (možno použít znak |)
brouk -volkswagen -vw - 2. a 3. termín nesmí být ve vyhledaných stránkách obsažen
notebooky filetype:xls - omezení jen na určitý formát dokumentu (zde XLS tabulka)
intitle:medicentrum interna - najde „medicentrum“ v názvu stránky a „interna“ kdekoli na stránce
allintitle:letový řád - najde termín „letový“ a zároveň „řád“ v názvu stránky
inurl:shop karajan - najde slovo „shop“ v URL a „karajan“ kdekoli na stránce
allinurl:search dvd - najde slovo „search“ a zároveň „dvd“ v URL
inanchor:dialog databáze - najde slovo „dialog“ v odkazu a „databáze“ kdekoli na stránce
allinanchor:digitální knihovna - najde slovo „digitální“ a zároveň „knihovna“ v odkazu
školení site:stk.cz - omezí na doménu „stk.cz“ a najde kdekoli na stránce „školení“
link:www.islandklub.com - najde stránky s odkazem na stránky Klubu islandských fanatiků
related:www.vlada.cz - najde stránky podobné stránce Úřadu vlády ČR
info:www.learning.cz - zobrazí informace o dané stránce (cache, related, link, +)
cache:www.mlp.cz spořilov - zobrazí danou stránku z cache a zvýrazní slovo „spořilov“
~help “excel 2002“ - zobrazí stránky o Excelu 2002 obsahující synonyma slova „help“

Pokud bychom se podívali na pokročilé (rozšířené) hledání obrázků, budeme již kromě tří kolonek na této stránce znát z pokročilého vyhledávání všechny ostatní. Je zde totiž možno zvolit velikost hledaného obrázku (libovolně velký, malý, střední, velký), jeho typ (jibovolný, JPG, GIF, PNG) a zabarvení (libovolné, černobílé, šedé, barevné).

Pomocí Googlu si můžete nechat překládat nějaký text nebo celou webovou stránku mezi různými jazyky. Čeština mezi nimi bohužel není, ale i tak může být tato pomůcka užitečná. Nachází se na stránce http://www.google.com/language_tools?hl=en

Dostáváme se k otázce bezpečnosti. Některým lidem (zvláště webmasterům, případně vedení firem) se u Googlu nelíbí funkce zobrazování stránek v podobě, jakou měly při návštěvě robota Googlu neboli Google Cache. Takovou funkci v širším měřítku nalezneme ještě např. u známého serveru Internet Archive a její odpůrci namítají, že když byly stránky z nějakého důvodu z internetu staženy, tak by neměly být někde stále dostupné. Sice existuje možnost, jak Googlu indexování stránek pro Google Cache zakázat (princip opt-out), ale je namítáno, že by spíše mělo být ručně povolováno (a implicitně zakázáno – princip opt-in). Tato funkce je však mnohými oblíbená kvůli možnosti objevit nějakou důležitou informaci na stránce z webu již stažené.

Nechtěl bych zde podávat přesný návod na to, jak je možno Googlu zneužít k neoprávněnému získání přístupu k databázím, prozrazení hesel, získávání nelegálního softwaru, Google-bombingu atd., ale myslím, že je dobré o podobných věcech vědět. Prostým zadáním vhodných termínů je možné najít stránky, které jsou administračním rozhraním do různých databází. Samozřejmě je pak odpovědnost především na administrátorovi konkrétních databází, aby je zabezpečil proti takovýmto průnikům. Jestliže někdo umístí na web stránku, na které bude nějaké heslo, musí počítat s tím, že ji může Google nalézt a heslo pak v něm bude vyhledatelné. Někdy je možné se setkat s hackery, kteří na web umísťují seznam uživatelských jmen a hesel, které při průniku do nějakého systému získali. Existuje velká pravděpodobnost, že se vám podaří pomocí Googlu nalézt nelegální software. Stačí vědět, že se pro něj často používá slovo warez a pro programy mající za úkol prolomit ochranu slovo crack. Pokud ještě vůbec existují programy, které pro svou registraci vyžadují pouze zadání sériového čísla, jsou existencí takovéhoto mocného vyhledávače rovněž ohroženy.

Google-bombing je termín, který byl v roce 2001 použit Adamem Mathesem, když se pokusil udělat si legraci ze svého kamaráda Andyho Pressmana a způsobil, že po zadání dotazu „talentless hack“ se v Googlu Pressmanova stránka Oh Messy Life objevovala na prvním místě. Fráze „talentless hack“ se nikde na jeho stránce nevyskytovala, ale to, že ji spolu s odkazem obsahovalo velké množství stránek jiných (hlavně weblogů po celém světě), k tomuto umístění stačilo. Podobným způsobem je možné doposud díky zneužití technologie Page Rank při zajištění dostatečného počtu odkazujících stránek (s vysokým hodnocením – Page Rankem) v seznamu výsledků posunout nějaké stránky na první místa. Háček je tedy pouze v tom, že stránky odkazující na tu, která se má v seznamu výsledků posunovat, musí mít také co nejvyšší Page Rank.

Pro server Google Watch je Google trnem v patě hned z několika důvodů: Jednak je to již zmiňovaná archivace stránek v systému Google Cache. Dalším důvodem je používání cookies, které mají platnost až do roku 2038. Každý, kdo navštíví Google, obdrží cookie (krátký textový soubor) do svého počítače a Google jej tak může při dalších návštěvách identifikovat. Pro Google je tedy pomocí cookie možné zjistit, co jste zde všechno hledali, kdy to přesně bylo, jaký jste použili prohlížeč nebo jakou jste měli IP adresu. Google k tomu dodává na stránce http://www.google.com/privacy.html, že zasílá cookies kvůli zlepšení kvality služby a lepšímu porozumění uživatelům. Neposkytne údajně získané informace třetí straně vyjma situace právního řízení jako je vydání zatykače, předvolání k soudu, zákonné opatření nebo soudní příkaz. Co se za tím skrývá doopravdy a k čemu jsou cookies zpracovávány, ví jen v Googlu. Dalším problémem je podle kritiků Googlu softwarový nástroj Google Toolbar, který si můžete stáhnout ze stránek Googlu a nainstalovat do Internet Exploreru. Měl by jako integrální součást prohlížeče urychlovat přístup k vyhledávání na Googlu, protože obsahuje vyhledávací políčko, které je tak při procházení různých stránek stále k dispozici, ale může také Googlu zasílat poslední vyhledávané termíny (opět za pomoci cookies). Navíc se aktualizuje automaticky bez upozornění.

Poslední velmi kritizovanou novinkou by mělo být spuštění nové služby Gmail. Sice má pro e-mailovou schránku poskytovat až 1 GB prostoru, ale zprávy by nemělo být možné mazat. Google by je podle vlastních kritérií měl časem promazávat sám. Ti, kteří o tuto službu projevili nebo projeví zájem vyplněním své mailové adresy na stránce About Gmail a nesmazali si předtím a potom cookies, mohou téměř jistě počítat s tím, že se tato zadaná adresa spáruje s daty o zadávaných dotazech, což nemusí být pro každého příjemné zjištění.

Bezpečeností a souvisejícími tématy u Googlu se zabývá např. Aleš Miklík ve svém dvoudílném článku Existuje odvrácená tvář Googlu? na Lupě.

Často se setkáváme s tvrzením, že když něco nebylo nalezeno v Googlu, tak to prostě nemůže na internetu být. Této představě by měli být informační specialisté vzdáleni. Jednak nemusí vyhledávat pouze na internetu, ale mohou využít i specializované databáze (přístupné přes internet či na CD-ROM) a jednak mohou využít i tradičních informačních zdrojů jako jsou slovníky, encyklopedie, tištěné bibliografie atd. Je třeba mít na mysli ověřitelnost údajů, které se dozvídáme ze stránek vyhledaných podobnými vyhledávači. Dále by informační specialisté měli aktivně vyhledávat v jiných vyhledávačích jako je například AltaVista, AlltheWeb, Ask Jeeves nebo v českém internetu Jyxo či Morfeo. Zkušenější se mohou pokoušet nalézt něco v oblasti „neviditelného webu“. „Jestliže by Google přestal fungovat, většina lidí by neměla plán B“, říká Joe Janes z Information School na University of Washington v Seattlu. „Knihovníci mají spoustu plánů B. Víme, kdy otevřít knihu, kdy někoho zavolat nebo i kdy použít pro hledání Google“ (článek When a Search Engine Isn't Enough, Call a Librarian online).

Google již není jen vyhledávač. Na stránce Google Services and Tools je možno se seznámit s dalšími službami a nástroji:

Froogle – vyhledávání cen různých produktů na webu

Google Answers – fórum, kde na rozličné otázky odpovídají za poplatek odborníci

Google Catalogs – možnost procházet tištěné reklamní katalogy

Google Groups – možnost komunikovat v diskusních skupinách Usenet

Google Labs – prototypy a projekty inženýrů Google ve vývoji

Google News – možnost prohledávat více než 4500 zpravodajských zdrojů

Google Special Searches – možnost vyhledávat na téma Apple, Microsoft atd.

Google University Search – vyhledává na stránkách konkrétní university

Google Directory – hierarchický katalog organizovaný do témat

Google Wireless – možnost přístupu ke Googlu z PDA zařízení

Blogger – možnost vytvořit si vlastní weblog

Google Browser Buttons – přidání tlačítek odkazujících na Google do prohlížeče

Google in Your Language – dobrovolný překlad vyhledávacího rozhraní

Google Toolbar – přidání vyhledávacího políčka do Internet Exploreru

Google Web APIs – nástroj pro použití Googlu určený programátorům