KAPITOLA 2

Jak pracují vyhledávače

V této kapitole:
- Jak pracují vyhledávače
- Jaké jsou technologie vyhledávání
- Jak se dívat na výsledky vyhledávání
- Co je SERP
- Vo znamená relevantní výsledek
- Jaké jsou typy vyhledávání
- Jaké vyhledávače využívat

Vědět, jak pracuje vyhledávač, je základní podmínkou úspěchu v internetovém marketingu. Vyhledávač je klíčovým nástrojem, který poskytuje odpovědi na dotazy uživatelů, a jako takový má významný vliv na to, co budou lidé číst nebo nakupovat.

Na Internetu je nepřeberné množství webových stránek, dokumentů, obrázků, zvukových nahrávek nebo videí. K tomu, aby je vyhledávač našel a poté je mohl nabídnout uživatelům, je nutné mít k nim přístup. Tento přístup je odkaz neboli link. Odborněji se nazývá hypertextový odkaz a je základním stavebním kamenem Internetu. Fakticky vzato, pokud na nějakou stránku (dokument, fotku apod.) nevede žádný odkaz, tak tato stránka neexistuje.

Poznámka: Odkaz na stránku nebo dokument musí být v klasickém HTML. V případě, že použijete JavaScript, tak jej robot neumí přečíst, a tudíž nebude brát odkaz v potaz.

Mohlo by se zdát, že vyhledávač (angl. – search engine) je pouze software procházející odkazy, které následně zobrazí jako výsledky na základě dotazu uživatele. Ve skutečnosti je to mnohem složitější. Vyhledávač představuje komplexní systém zahrnující nejen programy pro procházení stránek, ale také složité algoritmy pro řazení výsledků, databáze stránek a dokumentů, uživatelské rozhraní a také řadu informací získaných v průběhu používání vyhledávače. Vyhledávač si například pamatuje historii vyhledávání každého jednotlivého uživatele, pokud zná jeho identitu, nebo alespoň historii vyhledávání na každém použitém zařízení.

K tomu, aby mohl vyhledávač poskytnout uživateli odpověď na jeho dotaz, vede náročná cesta, byť je tato odpověď otázkou zlomku vteřiny. Práci vyhledávače lze zjednodušeně popsat ve 3 krocích:

Procházení (Crawling)

Nejprve program (software), kterému se říká robot (angl. – crawler nebo spider), prohledává stránky a sleduje odkazy v nich uvedené. Neznamená to však, že všechny stránky důkladně prozkoumává a zjišťuje, co na nich je. Zaznamenává v této chvíli jen informaci o tom, že stránka existuje. A to ještě ne o všech stránkách. Stránky prohledává robot pomocí odkazů. Takže pokud nevede na stránku žádný odkaz, robot ji technicky nemůže nikdy vidět. Toto procházení webem provádí robot stále. Na některé stránky se vrací častěji, na jiné méně často a další následně zcela opomíjí.

Indexování (Indexing)

Další fáze je indexace. Robot při indexaci vybírá klíčová slova, aby zjistil, o čem stránky jsou. Podle toho stránky třídí a pak je uloží do databáze, aby je mohl zobrazit ve výsledcích. Znamená to, že když uživatel zadá dotaz, robot znovu stránky neprochází, ale poskytuje výsledky ze stránek, které má v indexu.

Hodnocení (Ranking)

Robot má na základě indexace pro určitou stránku přiřazena klíčová slova. Po zadání dotazu pak posuzuje, do jaké míry je daná stránka relevantní k požadovanému klíčovému slovu v dotazu. Relevance se zjišťuje pro každé hledání zvlášť. Výsledky pak závisí na celé řadě faktorů a uplatnění nejrůznějších kritérií. Počet vyhledaných stránek (odkazů) je zobrazen (většinou) nad nebo pod výsledky vyhledávání.

Poznámka: Na Internetu jsou miliardy stránek, dokumentů, obrázků nebo videí, což představuje obrovské množství dat. Vyhledávače to vše musí mít k dispozici, to znamená, že musí tato data někam ukládat. K tomu slouží rozsáhlá datacentra rozmístěná (v případě Google) po celém světě, kde jsou umístěny desítky tisíc zařízení (počítače, disková pole atd.).

Relevantní výsledek vyhledávání

Uživatel zadá do vyhledávacího pole dotaz (slovo nebo slovní spojení) a dostane odpověď ve formě odkazů na stránky, dokumenty, obrázky, zvukové soubory, videa. Tyto výsledky jsou označovány anglickou zkratkou SERP – Search Engine Results Page. Řazení jednotlivých odkazů je uspořádáno tak, aby byl uživatel spokojen. Výsledek je tedy k jeho dotazu relevantní, tj. odpovídá zadanému dotazu.

Pro lepší orientaci v dalším textu je nutné upřesnit pojmy dotaz a klíčové slovo. Dotaz je řada slov zapsaných do řádku pro vyhledávání v nějakém vyhledávači. Tato slova se nazývají klíčová, avšak pro zjednodušení se používá jednotné číslo, tj. *klíčové slovo*. Pod tímto pojmem tedy rozumíme jak jedno slovo, tak více slov (většinou 2–5).

Poznámka: Podle průzkumů chování uživatelů se trend v rozšiřování počtu slov v dotazech zastavil, resp. obrátil. Zatímco v letech 2008–2009 výrazně rostly víceslovné dotazy, v letech 2011–2012 již tomu bylo naopak. Teorie dlouhého chvostu tak utrpěla určité trhliny.

Tabulka 2.1 Počet slov v dotazech – meziroční změny 2008–2009

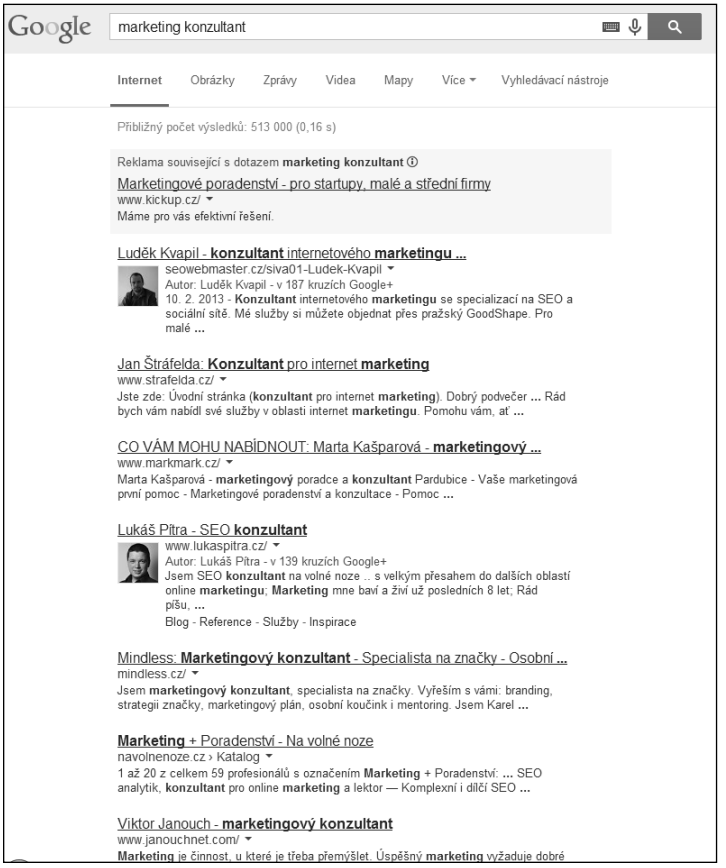
Procenta prokliků podle počtu slov (USA)				
Počet slov	**Leden 2008**	**Prosinec 2008**	**Leden 2009**	**Meziroční změna**
1 slovo	20,96 %	20,70 %	20,29 %	-3 %
2 slova	24,91 %	24,13 %	23,65 %	-5 %
3 slova	22,03 %	21,94 %	21,92 %	0 %
4 slova	14,54 %	14,67 %	14,89 %	2 %
5 slov	8,20 %	8,37 %	8,68 %	6 %
6 slov	4,32 %	4,47 %	4,65 %	8 %
7 slov	2,23 %	2,40 %	2,49 %	12 %
8 slov	2,81 %	3,31 %	3,43 %	22 %
Zdroj: Hitwise, an Experian company				

Tabulka 2.2 Počet slov v dotazech – meziroční změny 2011–2012

Procenta prokliků podle počtu slov (USA)				
Počet slov	**Květen 2011**	**Duben 2012**	**Květen 2012**	**Meziroční změna**
1 slovo	25,80 %	28,63 %	29,93 %	**+19 %**
2 slova	24,13 %	23,46 %	23,60 %	-2 %
3 slova	19,52 %	18,79 %	18,58 %	-6 %
4 slova	12,96 %	12,44 %	11,94 %	-9 %
5 slov	7,64 %	7,27 %	6,93 %	-11 %
6 slov	4,21 %	3,96 %	3,79 %	-12 %
7 slov	2,32 %	2,19 %	2,10 %	-11 %
8 slov	3,41 %	3,26 %	3,13 %	-10 %
Zdroj: Hitwise, an Experian company				

Poznámka: Na obrázku 2.1, resp. 2.2, je výsledek vyhledávání (SERP) pro dotaz „marketing konzultant". Google našel 513 000 výsledků, Seznam více než 545 000.

K tomu, aby se zobrazily relevantní výsledky vyhledávání, vede poměrně složitá cesta. Vyhledávače se snaží poskytnout výsledek takový, který uspokojí tazatele – odpovídá zadanému dotazu – a k tomu potřebují náročné technologie. Pracují nejen s faktory a jejich váhou, ale také s rozhodovacími stromy, technologiemi umožňujícími vzít v úvahu aktuální trendy, situaci na trhu, události v politice, sportu, kultuře, živelné katastrofy, epidemie a cokoliv dalšího vás napadne. Stejně tak berou v úvahu, o čem si lidé povídají na sociálních sítích, jaká videa sledují, co právě teď píší na Twitter, na co klikají apod. Následkem toho se mohou výsledky podstatně měnit. Na obrázku 2.3 je výsledek vyhledávání slova „biatlon" pár dní před zahájením olympiády v Soči. Na obrázku 2.4 pak 2 měsíce po jejím skončení.



Obrázek 2.1 Výsledky vyhledávání slova „marketing konzultant" pro Google

Uživatelé Internetu hledají nejrůznější věci a používají k tomu různé postupy, avšak obecně lze říci, že existují tři hlavní kategorie vyhledávání:

1. navigační vyhledávání (navigational search)
2. transakční vyhledávání (transactional search)
3. informativní vyhledávání (informational search)

Navigační vyhledávání znamená hledání nějaké URL nebo domovské stránky známé služby.

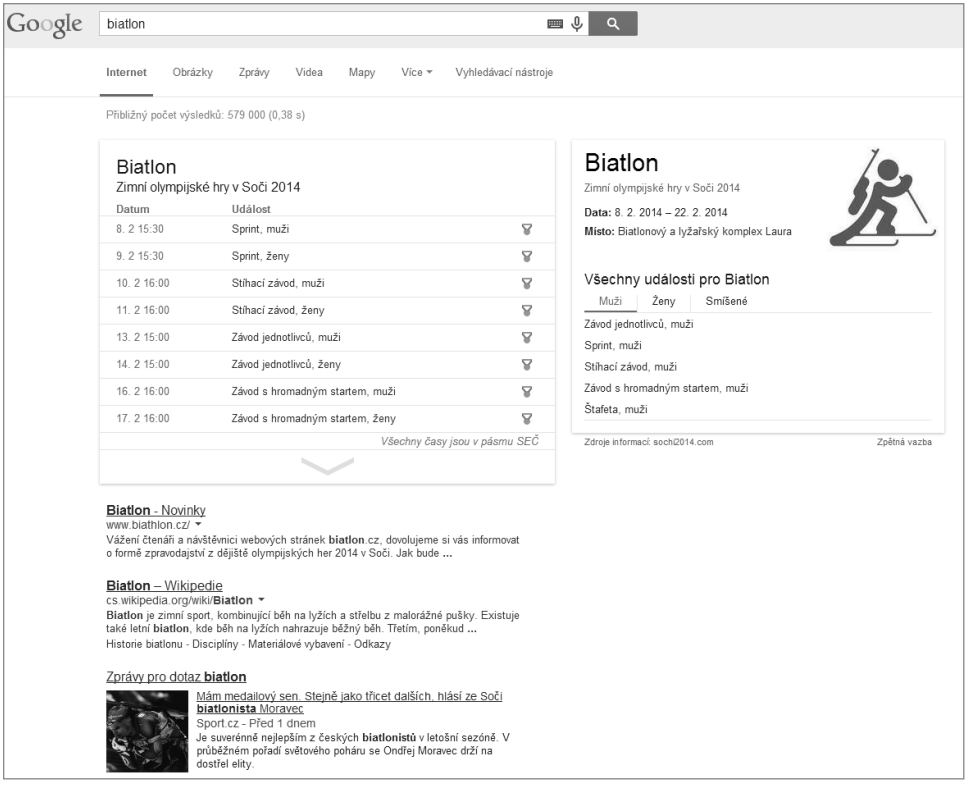
Transakční vyhledávání je obvykle vyhledávání za účelem nákupu. Například chcete koupit ubytování v hotelu. Pokud nemáte svůj oblíbený portál, kde budete ubytování hledat, tak pravděpodobně nejprve zadáte do vyhledávače slovo „hotel" a k tomu město, kde se chcete ubytovat. Pokud vás výsledky neuspokojí, budete pokračovat ve zpřesňování až do chvíle, kdy najdete hotel přesně podle vašich požadavků.

Obrázek 2.2 Výsledky vyhledávání slova „marketing konzultant" pro Seznam.cz

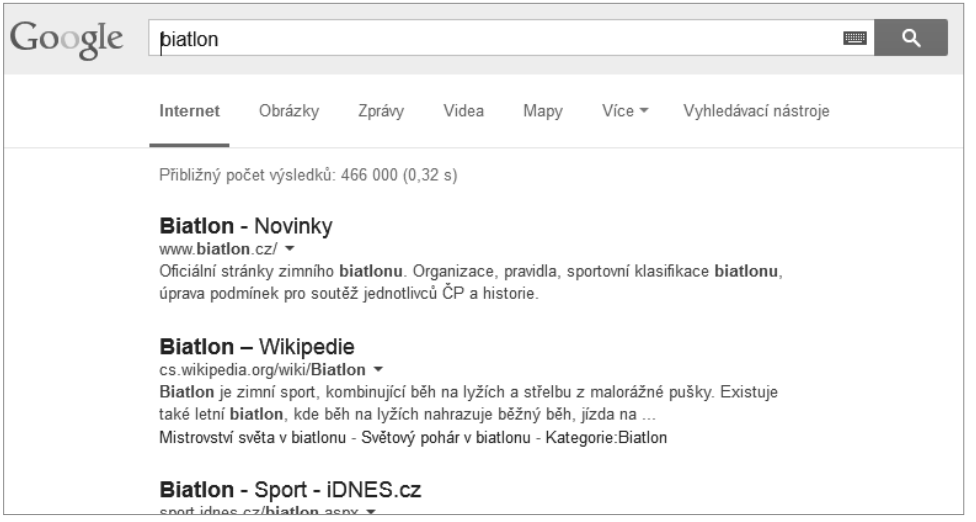
Informativní vyhledávání je například hledání názvu nějaké knihy nebo nejlepší restaurace ve městě.

Uvědomit si, že lidé postupují jinak při nákupu zboží a jinak třeba při hledání úryvku z knížky, je důležité pro vytváření webových stránek a e-shopů. Této problematice bude věnována kapitola 8.

Tip: Často se stává, že lidé nedostanou uspokojivou odpověď na svoji otázku. Může to samozřejmě být vinou vyhledávače, ale spíše je otázka špatně položena. Vyhledávače stále neumí (až na malé výjimky) odpovědět na otázku, ale hledají výsledky odpovídající slovům obsaženým v dotazu. Proto je třeba otázku zpřesňovat nebo se zeptat jinak. Zkuste najít přibližnou odpověď na otázku sami a tu pak napište jako dotaz. Ve většině případů dostanete relevantní výsledky.



Obrázek 2.3 Výsledky vyhledávání slova „biatlon" na začátku února 2014



Obrázek 2.4 Výsledky vyhledávání slova „biatlon" v červenci 2014

Technologie vyhledávání

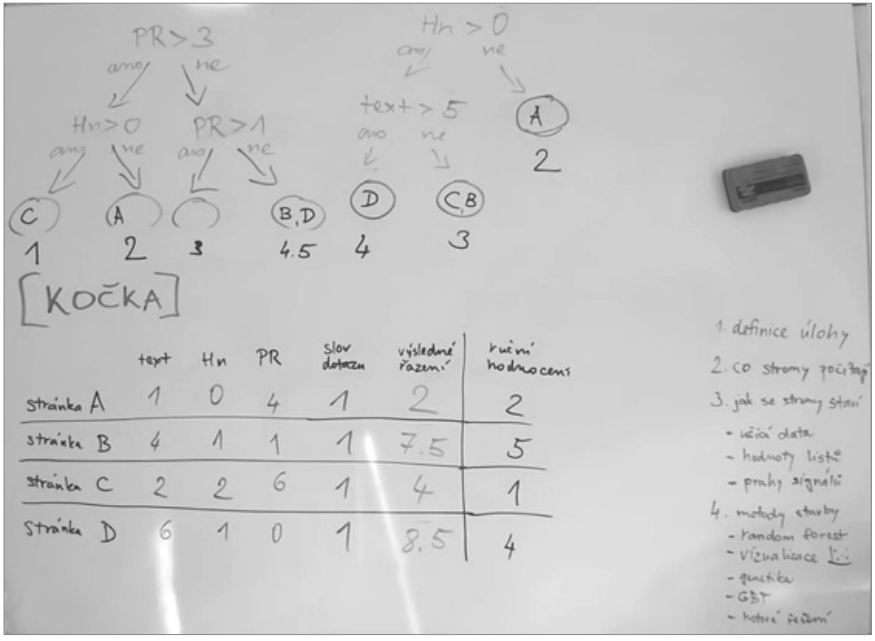
Přestože algoritmus vyhledávání je ostře střeženým tajemstvím každé firmy (vyhledávače), jsou známy některé technologie, jimiž se vyhledávání realizuje. Měli byste je znát, protože ovlivňují jak tvorbu webových stránek, tak marketingovou komunikaci.

Základem jsou faktory, podle nichž vyhledávač posuzuje kvalitu obsahu. Faktorů může být až 200, jako například klíčové slovo v titulku stránky, klíčové slovo v nadpisu, URL stránky, počet odkazů, zda jsou na stránce synonyma apod. Každý z těchto faktorů má svoji váhu, která se mění. Podrobněji o faktorech pojednává kapitola 8.

Různé speciální technologie se uplatňují jen v určitých případech, kdy je pro vyhledávač buď obtížné pochopit dotaz nebo si to vyžadují okolnosti. Je zřejmý rozdíl mezi sekvenčním hledáním (nevím přesně, co hledám) a hledáním strukturálním (vím, co chci, a postupně to upřesňuji). I toto musí vyhledávač umět rozpoznat a zvolit správnou technologii pro poskytnutí relevantního výsledku.

Regresní rozhodovací stromy

Regresní rozhodovací stromy pomáhají k řazení výsledků. Vyhledávač již tedy nějaké výsledky má, ale potřebuje vytvořit pořadí, tj. seřadit stránky podle relevance. K tomu využívá určité signály (například klíčové slovo v titulku, nadpisu, na stránce) a podle nich vypočítává určitou hodnotu, na jejímž základě sestavuje pořadí.



Obrázek 2.5 Příklad regresního rozhodovacího stromu (*Zdroj: http://youtu.be/3QLmEAIFZoA?list=PL932F62EECCA70D5E*)

Rozhodovacích stromů může vyhledávač používat velmi mnoho (desítky i více) a každý strom může být různě „košatý". V jednom stromu jsou například jen 3 úrovně, v jiném 10. Na obrázku 2.5 je příklad regresního rozhodovacího stromu.

Tip: Problematiku regresních rozhodovacích stromů popisuje velmi podrobně Dušan Janovský ve video seriálu, jehož 1. díl najdete zde: *http://www.youtube.com/watch?v=eXFWcbxAsSk&list=PL-932F62EECCA70D5E* (vhodné pro začátečníky i profesionály).

Lemmatizace

Další technologií, kterou využívají vyhledávače, je lemmatizace. Je to lingvistický nástroj na hledání základních tvarů slov. Lemmata = slova v základním tvaru. Tento nástroj používají všechny vyhledávače pro zjišťování, na co se vlastně uživatel ptá. Cílem je vyloučení nejednotnosti (desambiguace), tj. zvolení správného lemmatu.

Z pohledu vyhledávače je tedy důležité pochopit dotaz a poskytnout relevantní odpověď. Například někdo hledá „nabídka kurzů" a algoritmus rozpozná, že slovo „kurzy" v hledaném výrazu nemá souvislost s měnovými kurzy, ale se vzděláváním. V SERP to pak zohlední a nabídne stránky vztahující se ke vzdělávání, školení, seminářům. To znamená, že nabídne *nejlepší výsledek hledání = relevantní* (mající rozhodující význam, závažný, významný, důležitý, rozhodující – viz slovník cizích slov).

Stemming

Podobně jako lemmatizace funguje také stemming. Stemming je proces, při kterém se vyhledávače snaží rozpoznat skloňované slovo a vytvořit z něj infinitiv. V praxi hledají uživatelé slova v různých variantách a slovních spojeních. V momentě, kdy uživatel zadá slovo do vyhledávače, tak obvykle toto slovo nemá tvar infinitivu, ale je skloňované. Vyhledávač pak ve většině případů nabídne jen výsledky zobrazující infinitiv. Pro stemming existují různé typy algoritmů a každý má jinou funkci.

LDA

LDA (angl. – Latent Dirichlet Allocation) je technologie, která se snaží najít vysvětlení, proč jsou si některé části dat podobné. Pomocí LDA lze zjistit v textu určité skupiny témat a každé slovo pak přiřadit k některému z nich. Tato technologie je pravděpodobně používána při vyhledávání v případech, kdy neexistují zpětné odkazy na obsah (nebo nejsou relevantní). Podle provedených testů se zdá, že v případech hledání málo konkurenčních slov s dlouhým chvostem tato technologie skutečně funguje. Naopak u konkurenčních a kratších slov nemá LDA téměř žádnou váhu (podrobnosti v angl. na: *http://www.seomoz.org/blog/lda-and-googles-rankings-well-correlated*).

QDF

QDF (Query Deserves Freshness) má vliv na změny v SERP s ohledem na aktuální události a/nebo frekventované dotazy. Vyhledávané fráze se mění poměrně často vlivem nejrůznějších sportovních, politických nebo jiných událostí. Kromě toho dochází mezi lidmi k šíření různých slov,

protože je používají známé osobnosti například v televizi. Dochází tak nejen k pokládání nových dotazů, ale i ke změně frekvence určitých frází.

Poznámka: Google i Seznam vydávají pravidelně žebříčky nejhledanějších frází. V tabulce 2.3 je žebříček od Google, tzv. Duch doby, v tabulce 2.4 pak od Seznam.cz. Pokud se jako marketéři chcete svézt na vlně zájmu, neměli byste je přehlížet. Na druhou stranu to samé asi udělají vaši konkurenti, takže se spíše soustřeďte na dlouhý chvost.

Tabulka 2.3 Nejhledanější slova na Google za rok 2013

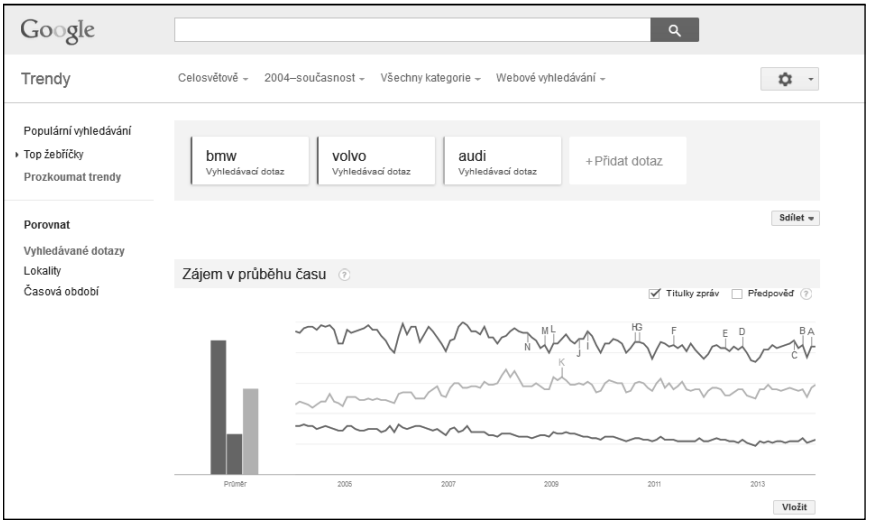
Nejhledanější výrazy	Skokani roku	Nejhledanější zdroje informací	Nejhledanější výrazy ve spojení s „Jak“	Nejhledanější výrazy ve spojení „Co je“
Facebook	Vyvolení	IDOS	Jak zhubnout	Co je tablet
YouTube	Karel Schwarzenberg	TV program	Jak sbalit holku	Co je NFC
Překladač	Volby 2013	Počasí	Jak uvázat kravatu	Co je Instagram
iDnes	MS hokej 2013	Wikipedie	Jak naladit Smíchov	Co je Swag
IDOS	Google Play	ČSFD	Jak psát web	Co je internet
Mapy	Velikonoce 2013	Heuréka	Jak vydělat peníze	Co je Twitter
Aukro	Miloš Zeman	ARES	Jak zrušit Facebook	Co je Linkedin
Blesk	Iveta Bartošová	Mapy	Jak napsat životopis	Co je IBAN
Gmail	ČT Sport	Slovník	Jak sbalit kluka	Co je PayPal
TV program	Superstar 2013	Justice	Jak naladit Déčko	Co je kuskus

Tabulka 2.4 Nejhledanější slova na Seznam.cz za rok 2013

Nejhledanější slova	Skokani mezi událostmi a kauzami	Televizní a filmoví skokani	Skokani mezi komerčními produkty	Skokani mezi kulturními a sportovními akcemi
Facebook	Povodně	Česko Slovensko má talent 2013	YouTube	Velká pardubická 2013
YouTube	Volby	TN.cz zprávy	Klimatizace do bytu	Barum Rally 2013
Google	Zatmění Měsíce	Superstar 2013 kdo vypadl	Obezin	Miss 2013
Porno	Volby prezidenta ČR	Vyvolení online	E-dražby	TýTý 2013
Freevideo	Petice proti amnestii	Ordinace v růžové zahradě poslední díl přehrát zdarma	Ask.fm	Český slavík 2013
Ulož.to	Paterčata v Česku	Vařte jako šéf recepty	Kanadská krbová kamna	Davis Cup
Bazoš	Anastasia Hagen	Výměna manželek 30. 8. 2013	Svatební šaty DD	Ples v Opeře 2013
Superhry	Zeman a korunovační klenoty	Ve stínu	Manukový med	Český lev 2013

Nejhledanější slova	Skokani mezi událostmi a kauzami	Televizní a filmoví skokani	Skokani mezi komerčními produkty	Skokani mezi kulturními a sportovními akcemi
Aukro	Titanic Palace	Rychle a zběsile	Xiaomi Mi2S	Dakar 2013 online
Idnes	Kate Middleton porodila	Habermannův mlýn	Dárkové sady kvalitní kravaty	Noc kostelů 2013

Tip: Pokud chcete zjistit frekvenci pokládaných dotazů na Seznam.cz, pak stačí vždy kliknout dole na stránce výsledků vyhledávání na odkaz Statistika hledanosti. Google má obdobnou službu (Google Trends: *www.google.com/trends/*), která však neukazuje konkrétní počty hledání určitého výrazu. Ukazuje relativní hodnoty za určité období, ale navíc umožňuje porovnávání několika výrazů. Na obrázku 2.6 je vidět poměr hledání slov bmw, volvo a audi.



Obrázek 2.6 Výsledky porovnání hledaných slov v Google Trends

QDD

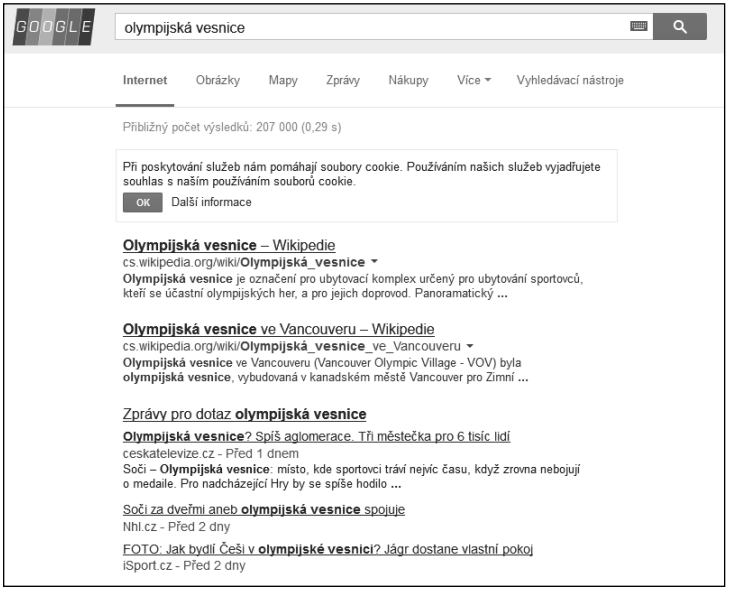
QDD (Query Deserves Diversity) se používá při podobných nebo nejednoznačných výsledcích na určitý dotaz, kdy vyhledávač bere v úvahu příbuzné fráze.

V případě, že na hledaný dotaz jsou výsledky (nalezené stránky) velmi podobné nebo je vyhledávací dotaz mnohoznačný, bere algoritmus v úvahu ještě příbuzná klíčová slova, a teprve když zjistí výskyt i těchto klíčových slov na dané stránce, nabídne ji ve výsledcích. Dochází tak k tomu, že obsahově bohatší stránka má přednost před stránkou optimalizovanou jen na určitou frázi (klíčové slovo). Příbuzné fráze může například zjistit z historie posloupnosti dotazů.

Vyhledávání v reálném čase

Při položení určitého dotazu může vyhledávač usoudit, že se uživatel ptá na něco aktuálního. Pak bude brát v potaz aktuální zdroje informací více než ostatní. Jako aktuální zdroj informací slouží

zpravodajské portály, Twitter, diskusní fóra apod. Na obrázku 2.6 je vidět výsledek vyhledávání dotazu „olympijská vesnice“, kde jsou také úryvky z aktuálního zpravodajství.



Obrázek 2.7 Výsledky vyhledávání dotazu „olympijská vesnice“

Citační analýza

Jedná se o prověřování externích linek vedoucích na stránky. Tato analýza se dělá vždy v kontextu dotazu! Hodnotí se i kvalita stránek s odkazem. Význam má i umístění odkazu (v textu má větší význam než např. někde dole na stránce nebo v menu). Citační analýza je klíčovým prvkem pro řazení odkazů. Vyhledávače se správně domnívají, že počet a kvalita odkazů znamená určitou popularitu dané stránky, a proto takovou stránku posouvají výše ve výsledcích vyhledávání. Je třeba zdůraznit, že jde hlavně o kvalitu odkazu. Velké množství odkazů (např. z katalogů) nemá vůbec žádný význam a vyhledávače takové odkazy ignorují.

Na druhou stranu se může na Internetu nacházet velmi zajímavý obsah, na nějž nikdo neodkazuje. V takovém případě se vyhledávač zaměřuje především na analýzu obsahu.

Sémantické vyhledávání

Sémantické vyhledávání znamená hledání podle obsahu. Výhradně na tomto principu již fungují některé speciální vyhledávače (jako Hakia), ale i Google tento princip používá a podle vyhledávacího dotazu uplatňuje některé zmíněné technologie.

Sémantické vyhledávání nebere v úvahu tzv. popularitu stránky (počet a kvalita odkazů na konkrétní stránku), ale již zmíněný obsah (v některých případech jde o kombinaci obou metod). Stránky ve výsledcích pak nemusí nutně obsahovat vyhledávací dotaz, ale například synonyma. Budete například hledat slovní spojení „domácí čistička vzduchu“, ale na stránkách se bude psát

o „klimatizaci do bytu". Pokud vyhledávač usoudí, že tato stránka má vysokou relevanci, může ji nabídnout na lepší pozici ve výsledcích než stránky obsahující zadané slovní spojení.

Normalized Google Distance

Na základě tohoto teoretického konceptu vytvořil Josef Šlerka webovou aplikaci Mechanical Cinderella (*www.mechanicalcinderella.com*). Základem je měření podobnosti mezi slovy pomocí internetových vyhledávačů. Šlerka ve své práci ukazuje, jak lze zkoumat kolektivní inteligenci v oblasti jazykové sémantiky a pragmatiky s využitím klasických přístupů informační vědy a současných vyhledávačů.

Systém je založen na počítání sémantické blízkosti slov. Teoreticky co je blízké, patří do společné kategorie. V češtině bude pravděpodobně blízké slovo ke slovu dovolená Itálie, naopak Etiopie asi sotva. Princip je vcelku jednoduchý – počítá s výskytem slov na Internetu. Z toho pak vychází blízkost slov.

V tabulce 2.5 je vidět, že u slov blízkých se číslo blíží nule a naopak. Z tabulky vyplývá, že nejsilnější vazba je mezi slovy dovolená a Německo. Samozřejmě to neznamená, že nejvíce lidí jezdí na dovolenou do Německa, ale že mezi těmito slovy existuje bližší vztah než třeba mezi dovolenou a Kolumbií. Když pak budeme zkoumat záplavy, zjistíme jen minimální souvislost s Kolumbií.

Tabulka 2.5 Výsledky hledání sémanticky blízkých slov

	Itálie	Rusko	Německo	Kolumbie
dovolená	0,25452514	0,26195551	0,22943152	0,30356721
bezpečnost	0,33035133	0,35602984	0,3875683	0,39261453
fotbal	0,32008766	0,41883921	0,34105614	0,44020754
záplavy	0,31320026	0,30324527	0,36001687	0,83192649

Marketéři mohou tyto výpočty využít například při plánování cílení na určitá klíčová slova.

V souvislosti se sémantickým vyhledáváním se tak dostává do popředí kvalita obsahu, která má pro vyhledávání klíčový význam. Pochopení obsahu je pro vyhledávače velice obtížné, ale jsou v tom stále úspěšnější. K tomu přispívá také to, že Google má v indexu velké množství knih, z nichž může čerpat informace a učit se, jak vypadá přirozený obsah. Právě text literárních děl obsahuje zásadní faktory pro identifikaci jazyka, protože neobsahuje žádná reklamní sdělení nebo oborovou hantýrku.

Vyhledávání obrázků

Pro vyhledávání obrázků používají vyhledávače jednak alternativní text v HTML kódu (alt) a také pokročilejší technologie umožňující zjišťovat, co se na obrázku nachází. Google má k dispozici jednak nástroj pro rozpoznávání textu, který je v obrázku, a ze zkušenosti jistě víte, že je schopen také rozpoznávat barvy. K dispozici jsou rovněž technologie pro rozpoznávání tváří. Nejen Google již tyto technologie používá a předpokládá se, že se ještě v dokonalejší formě uplatní u chytrých brýlí (Google Glass, Recon Jet).