VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

PROBLEMATIKA SEO SPAMU A METODY JEHO DETEKCE

DIPLOMOVÁ PRÁCE
MASTER‘S THESIS

AUTOR PRÁCE BC. KAREL TLUSŤÁK
AUTHOR

BRNO 2010

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

PROBLEMATIKA SEO SPAMU A METODY JEHO DETEKCE

THE SEO SPAM AND METHODS OF DETECTION

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — BC. KAREL TLUSŤÁK
AUTHOR

VEDOUCÍ PRÁCE — ING. ZBYNĚK MICHLOVSKÝ
SUPERVISOR

BRNO 2009

Abstrakt

Tato práce se zabývá problematikou vyhledávání ve fulltextových vyhledávačích a možnostech ovlivňování výsledků vyhledávání pomocí Search Engine Optimalization spamu. Jsou zde popsány principy vyhledávačů, způsoby ovlivňování pomocí SEO a spamu. Dále jsou zde popsány současné metody, které byly navrženy proti tomuto ovlivňování a jejíž snahou je zkvalitnit relevantnost výsledků vyhledávání. Součástí této práce je rovněž návrh vlastní metody, její implementace a testování úspěšnosti nad vytvořenými testovacími daty.

Abstract

This paper is concerned in the full-text searching problematic and possibilities of influencing search result pages by mean of Search Engine Optimalization spam. Principles of search engines and methods which are used for SEO spam are described here. This work also shows actual methods which are used to reduce this spam effect and were proposed to improve relevancy of search results. Part of this paper is design of new method against SEO spam, its implementation and testing performance of this method upon created testing data.

Klíčová slova

Seo spam, black hat seo, detekce seo spamu

Keywords

Seo spam, black hat seo, detekce seo spamu

Citace

Karel Tlusťák: Problematika SEO spamu a metody jeho detekce, diplomová práce, Brno, FIT VUT v Brně, 2010

Problematika SEO spamu a metody jeho detekce

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením pana Ing. Zbyňka Michlovského.
Další informace mi poskytli Mgr. Jiří Vraný, Ph.D.
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

……
Karel Tlusťák
17.5.2010

Poděkování

Rád bych tímto poděkoval za velmi produktivní spolupráci vedoucímu této diplomové práce panu Ing. Zbyňku Michlovskému, dále pak za odbornou konzultaci, kterou mi poskytnul pan Mgr. Jiří Vraný, Ph.D. a rovněž za čas, ochotu a cenné rady panu Ing. Dušanu Janovskému.

© Karel Tlusťák, 2010

Tato práce vznikla jako školní dílo na Vysokém učení technickém v Brně, Fakultě informačních technologií. Práce je chráněna autorským zákonem a její užití bez udělení oprávnění autorem je nezákonné, s výjimkou zákonem definovaných případů.

Obsah

Obsah 1
1 Úvod 3
2 Principy vyhledávačů a SEO 4
2.1 Vyhledávače, jejich rozdělení a funkce 4
2.2 Metriky pro sestavení SERP 5
2.2.1 Určení obsahové hodnoty stránek 6
2.2.2 Důležitost stránky 7
2.3 Search Engine Optimalization 9
2.3.1 SEO z pohledu internetového marketingu 9
2.3.2 Etické metody SEO 11
2.3.3 Zakázané metody SEO 12
2.3.4 Definice SEO spamu 16
3 Metody obrany proti SEO spamu 16
3.1 Všeobecné požadavky na kvalitu vyhledávání 17
3.1.1 Problém určení autoritativních domén 17
3.2 Metody detekce SEO spamu 22
3.2.1 Známé metody využívající analýzu hypertextových odkazů 22
3.2.2 Metody využívající analýzu on-page faktorů 30
3.3 Budoucnost detekce SEO spamu 33
4 Měření kvality výsledků vyhledávání 34
4.1 Tvorba kalibračních dat 36
4.2 Použití kalibračních dat 36
4.2.1 Automatické nastavování parametrů vyhledávače pomocí kalibračních dat 37
5 Návrh vlastní metody proti SEO spamu 37
5.1 Princip navrhované metody 38
5.2 Návrh způsobu výpočtu velikosti množin m 39
5.3 Ukázka výpočtu velikosti množin m 40
5.4 Možné modifikace a vylepšení způsobu výpočtu velikosti množiny 43
5.5 Určení míry SEO spamu pomocí metody korelace výsledků vyhledávání 44
5.5.1 Metoda rozdílu SERPů 44
5.5.2 Metoda vztahu trp 45
5.5.3 Metody analýzy více faktorů 45
6 Implementace metody korelace výsledků vyhledávání 46
6.1 Implementační prostředí 46

6.2 Uživatelské rozhraní 47
6.3 Paralelizace zpracování 48
7 Tvorba kalibračních dat 49
7.1 Průměrné ohodnocení jednotlivý stránek výsledků vyhledávání 50
8 Testování úspěšnosti implementované metody 56
8.1 Srovnání výsledků, úspěšnost metody 57
9 Závěr 58
Literatura 59

1 Úvod

V současné době je velmi aktuální potřeba rychle a efektivně hledat informace. K tomuto používáme především internetové vyhledávače, které nám zobrazí odkazy na nalezené stránky, se kterými již dále pracujeme. Je tedy velmi důležité, aby zobrazené výsledky byly relevantní a ušetřili náš čas hledání. Snahou všech vyhledávačů je tedy zobrazovat co nejrelevantnější výsledky hledání vzhledem k danému dotazu. Na druhou stranu z pohledu provozovatelů samotných webových stránek (obzvláště komerčního charakteru) je snahou „být vidět“ – tedy ovlivnit mínění internetového vyhledávače o relevantnosti dané stránky na uvažovaný dotaz. K tomuto se dá použít tzv. SEO spam – jedná se o metody, jejichž snahou je uměle zvýšit důležitost stránky a tím zlepšit pozice ve vyhledávání. Mezi tyto metody patří například umělý linkbuilding (nakupování zpětných odkazů), generování mikrostránek, doorway stránky, cloaking, skrytý text, atd.

Fenomén SEO spamu representuje hlavní problém většiny internetových vyhledávačů [1]. Problém SEO spamu se nedotýká jenom vývojářů samotných internetových vyhledávačů, ale i celé internetové komunity, jelikož negativně ovlivňuje rychlost a přesnost vyhledávání informací na tomto médiu.

Tato práce se snaží analyzovat, jaké metody SEO spameři používají a jak jsou jednotlivé metody účinné. Dále pak shrnuje, jaké jsou způsoby obrany proti SEO spamu, a navrhuje vlastní metody pro zkvalitnění vyhledávání a boje proti SEO spamu. Součástí této práce je i samotná implementace jedné z navrhovaných metod. Dále je sestaven vzorek tzv. kalibračních dat – výsledků vyhledávání a jejich relevanci k danému hledanému výrazu. Na tomto vzorku dat je následně testována úspěšnost implementované metody.

Existuje mnoho metod, kterými se internetové vyhledávače snaží proti SEO spamu bránit, nicméně nelze z určitostí říci, které metody jsou reálně využívány a jak moc jsou efektivní či využitelné. Důvodem tohoto nedeterminismu je především fakt, že samotné společnosti provozující vyhledávače přísně své implementované metody tají. Uveřejnění takovýchto informací by totiž napomohlo SEO spamerům ke zdokonalování stávajících či vývoji nových metod.

2 Principy vyhledávačů a SEO

Tato kapitola popisuje principy vyhledávačů a nejčastější metriky, které vyhledávače používají pro hodnocení důležitosti stránek a sestavení seznamu nalezených webů vzhledem k danému dotazu (tzv. SERP – search engine result page). Dále se zabývá všeobecnými metodami, které se označují jako SEO (search engine optimalization) a shrnuje, které z těchto metod jsou označovány jako etické a které nikoliv. Na tomto základu je pak definováno, co budeme v rámci této práce považovat za SEO spam.

2.1 Vyhledávače, jejich rozdělení a funkce

Vyhledávače jsou hlavním pojítkem mezi uživatelem internetu a samotnými informacemi. Již od počátku internetové éry byla snaha usnadnit uživatelům přístup k informacím pomocí tzv. katalogů. Katalogové vyhledávání je nejstarší technikou třídění informačních kanálů. Vyhledávací robot v tomto případě disponuje uměle naplněnou databází subjektů roztříděnou podle mnoha vyhledávacích hledisek. Výhodou tohoto přístupu je fakt, že informace v databázi jsou většinou schvalovány či překontrolovány člověkem a lze tedy předpokládat, že jsou správné a ověřené. Další výhodou katalogového vyhledávání je fakt, že katalog může poskytovat velmi specifické informace, podle kterých lze vyhledávat a řadit (například množství zaměstnanců firem v katalogu hospodářské komory České Republiky). Velkou nevýhodou těchto vyhledávačů je však fakt, že jejich databáze je omezena pouze na entity, které byly do katalogu uměle vloženy (a zařazeny) a tyto vyhledávače je tedy třeba manuálně aktualizovat. Další nevýhodou je, že v katalogu je poměrně omezené množství informací, což omezuje možnosti samotného vyhledávání. Jedná se o to, že katalog již ze své podstaty nemůže obsáhnout celou šíři internetu. Proto se katalogové vyhledávače většinou úzce specializují na určité informační odvětví, v rámci kterého bude vyhledáváno. Existuje též skupina vyhledávačů, které jsou částečně katalogové a částečně fulltextové. Jedná se většinou o vyhledávače, jejichž zdroje dat jsou tzv. feedy (xml, rss, atd) a které vyhledávají částečně fulltextově a částečně podle parametrů uživatele (například uživatel chce zobrazit pouze zboží, které mají eshopy skladem).

V současné době převládá fulltextové vyhledávání. Důvodem tohoto trendu je fakt, že fulltextové vyhledávače jsou většinou schopné vyhledávat v rámci celého internetu (nebo alespoň jeho regionální části) a netrpí výše zmíněnými nevýhodami katalogových vyhledávačů. Tyto vyhledávače většinou používají robota, který se pomocí hypertextových odkazů pohybuje po internetu a indexuje do své databáze obsahy jednotlivých stránek a jejich provázanosti. Díky tomuto mechanismu je schopen udržovat aktualizovanou databázi obsahů stránek, mezi kterými pak vyhledává. Na druhou stranu je zde velkým problémem kvalita a relevantnost samotných výsledků

vyhledávání, jelikož informace v databázi jsou tak obsáhlé a často obměňované, že nemohou být ověřeny administrátory, jako tomu bývá u katalogových vyhledávačů.

2.2 Metriky pro sestavení SERP

Základním problémem všech vyhledávačů je samotné sestavení pořadí výsledků vyhledávání. Jedná se o tzv. SERP (Search Engine Result Page) – stránka obsahující nejčastěji několik webových odkazů o kterých si vyhledávač myslí, že obsahují právě tu informaci, kterou uživatel hledal. Tyto odkazy jsou vzestupně seřazeny podle několika metrik a obsahují krátké popisky, v nichž jsou zvýrazněna hledaná slova. Tato stránka může rovněž obsahovat placené reklamní sdělení či další prvky.

Každý vyhledávač používá pro sestavení SERP jiné metriky. Všeobecně lze říci, že katalogové a jim příbuzné vyhledávače používají buď jako metriku specifický parametr zadaný uživatelem (například počet zaměstnanců) nebo v případě nerozhodnosti řadí výsledky podle abecedy. Jelikož jsou katalogové vyhledávače většinou komerční, řazení je též ovlivněno bonitou společnosti zastupující danou webovou stránku. Často se lze setkat s modelem, kdy existují různé balíčky tzv. garantovaných zápisů a výsledky vyhledávání jsou ovlivněny tím, do které marketingové skupiny zápisu daná společnost patří. Tento model je patrný například u nejznámějšího českého katalogového vyhledávače firmy.cz. Tento vyhledávač obsahuje databázi katalogově zařazených firem a dle uživatelova dotazu nalezne nebližší relevantní kategorii a stránky. SERP tedy obsahuje výpis firem zařazených v nalezené nejrelevantnější kategorii a rovněž seznam dalších relevantních kategorií. Tento výpis firem je opět seřazen podle několika kritérií. První jsou vždy (bez ohledu na to, na jaké je uživatel stránce) tzv. sponzorované odkazy – jedná se o společnosti, které si zaplatili sponzorství dané kategorie. Dále jsou společnosti, které si koupili tzv. zápis s logem. Tyto firmy jsou seřazeny podle celkové bonity u společnosti Seznam a.s. Za těmito firmami jsou tzv. negarantované zápisy, čili ostatní firmy, které za pořadí katalogu firmy.cz nechtějí platit.

U některých částečně katalogových a částečně fulltextových vyhledávačů (např. srovnávače cen) se většinou používá pro sestavení SERP několik metrik. Jelikož každý vyhledávač má své vlastní odlišnosti, tyto metriky budou ukázány na modelovém případě reálného vyhledávače zbozi.cz. Tento vyhledávač tvoří SERP podle několika kritérií v závislosti na typu hledání, který si uživatel zvolil (Dle nejlepší shody, od nejlevnějšího / nejdražšího). Pokud si uživatel zvolí metodu „od nejlevnějšího“, SERP je vytvořen podle ceny jednotlivých položek (obdobně v případě „od nejdražšího“) zajímavější je z hlediska problematiky tvorby SERP metoda „Dle nejlepší shody“. V tomto případě jsou používány následující metriky

- Relevance k danému výrazu
- Komerční aspekty

Komerční aspekty zde znamenají, že stránky, ze kterých pochází vyhledávané zboží, se rozdělí na 2 skupiny – 1. skupina jsou stránky, jejichž majitel má se společností Seznam a.s. (provozující

srovnávač zbozi.cz) smluvní závazek placených prokliků, 2. skupina jsou ostatní neplacené stránky. Všechny nalezené produkty odpovídající hledanému výrazu ze stránek v 1. skupině mají vždy přednost před produkty ze stránek v 2. skupině. Relevance k danému výrazu je způsob, kterým portál zbozi.cz řadí produkty v dané skupině. Předpokládá se, že jsou zde použity podobné metriky řadění jako u fulltextového vyhledávání s tím rozdílem, že se přihlíží k charakteristice zdrojových dat (např. RSS feed). Tyto srovnávače se tedy dají částečně ovlivnit pomocí SEO spamu.

Z hlediska tvorby SERP je nejproblematičtější fulltextové vyhledávání. Většina fulltextových vyhledávačů se totiž snaží poskytovat jen relevantní výsledky vyhledávání, které nejsou ovlivněny komercí (nelze si tedy „zaplatit pozici“). Zjednodušeně lze tedy říci, že vyhledávače nejprve vyhledají stránky, na kterých se vyskytuje hledané slovní spojení a tyto stránky pak seřadí podle obsahové hodnoty stránky vzhledem k vyhledávanému spojení a podle důležitosti (důvěryhodnosti) samotné stránky. Z takto seřazených stránek je již schopen sestavit SERP a ten uchovávat v cache po určitou dobu (předpokládá se totiž, že spousta uživatelů zadává stejné dotazy). Problémem je zde však objektivní určení oné obsahové hodnoty a důležitosti. Toto je tedy hlavní bod zájmu nejen vyhledávačů ale zároveň i SEO spamerů.

2.2.1 Určení obsahové hodnoty stránek

Jelikož žádný vyhledávač neuveřejňuje, podle kterých kritérií přesně určuje obsahové hodnoty stránek, nelze jednoznačně popsat metodiku určování takovýchto hodnot. Všeobecně však platí, že vyhledávače při analýze stránky používají následující faktory, pomocí kterých lze posuzovat „čím se daná stránka zabývá“:

- Slova obsažená v titulku stránky (v tagu <title>)
- Slova obsažená v url dané stránky
- Slova sémanticky zvýrazněná (např. pomocí <H1>, <STRONG>, atd.)
- Výskyt slov v anchorech odkazů na stránce
- Densita slov v samotném obsahu stránky
- Rozložení výskytu slova v textu, podobná slova, synonyma
- Doménové faktory (stáří, doba předplacení, klíčová slova v názvu domény, atd.)
- Výskyt slov v popiscích obrázků (toto je používáno především pro vyhledávání obrázků)
- Výskyt slov v meta-informacích (keywords, description)

Analýzou těchto faktorů lze vypočítat, jak jsou jednotlivá slova / slovní spojení na stránce významná a kolik informací o kterých tématech daná stránka poskytuje. V rámci této práce jsem vyvinul analyzátor těchto faktorů, který bude později použit pro potřeby analyzování SEO spamu.

2.2.2 Důležitost stránky

Předchozí analýza zohledňuje pouze obsah samotných stránek (tzv. on-page faktory), ten však lze poměrně snadno měnit a je tedy obtížné objektivně posoudit důvěryhodnost tohoto obsahu a celkovou významnost stránek. Ideální by bylo posuzovat vlastní zkušenosti návštěvníků s jednotlivými stránkami. Z této filosofie vychází 2. nosný pilíř, pomocí kterého vyhledávače určují pořadí stránky v SERP. Většina vyhledávačů používá jako referenční model zkušeností návštěvníků a samotné oblíbenosti (tudíž obsahové bohatosti) stránky tzv. zpětné odkazy (back links). Jedná se o hypertextové odkaz směřující z cizích domén na uvažovanou stránku. Tento model vychází z předpokladů, že jestliže je stránka kvalitní a obsahově bohatá, lidé na ni upozorňují další uživatele internetu tím, že zveřejní její adresu (např. na svých stránkách, v rámci diskusního fóra atd). Jiným pohledem na tento model může být analogie referenční literatury (seznam citací či zdrojů) v knihách. Lze tedy říci, že čím více lidí se odkazuje na určitou stránku, tím je důležitost této stránky vyšší. Při tvorbě SERP je tedy nutné posuzovat nejen to, jak je stránka obsahově relevantní na požadované slovo, ale i kolik jiných stránek je o tomto faktu přesvědčeno. Moderní vyhledávače tedy nesledují pouze počty zpětných odkazů, ale hodnotí i ze kterých stránek tyto odkazy pocházejí (původ odkazů) a jaký je text samotných odkazů a jejich okolí.

2.2.2.1 Hodnocení původu odkazu

Při hodnocení zpětných odkazů vyhledávače sledují především to, jak důležitý je zdroj, který na naši stránku odkazuje. Čím důležitější je odkazující stránka, tím větší váhu má tento odkaz. Zde je však problém opětovného určení důležitosti odkazující stránky – ta musí být vypočtena opět pomocí analýzy zpětných odkazů směřující na tuto stránku, přičemž tyto odkazující stránky musí být rovněž analyzovány a takto bychom teoreticky mohli pokračovat do nekonečna. Jak bude vysvětleno později, tyto výpočty se počítají v iteracích, které se blíží požadované hodnotě zastupující onu „důležitost“. Moderní vyhledávače nepočítají jen kolik a jakých odkazů směřuje na uvažovanou stránku, ale berou v úvahu i další faktory těchto odkazů. Například dle jednoho z patentů google posuzuje stáří domén odkazujících stránek, rychlost nárůstů zpětných odkazů, dobu, po kterou stránka odkazuje, množství stránek, a další. [2]

2.2.2.2 Hodnocení textu odkazu a jeho okolí

Jelikož samotný fakt, že stránka odkazuje na jinou stránku, nemá vypovídající hodnotu o sémantice tohoto odkazu, moderní vyhledávače se snaží rovněž analyzovat, jaké jsou další parametry takovéhoto odkazu. Při této analýze hrají klíčovou roli tzv. anchory odkazů a okolí odkazu. Anchorem se v tomto případě myslí textová reprezentace odkazu, kterou vidí uživatel a která de facto uživateli napovídá, co odkaz nabízí za informace. Uvažujme následující odkaz:

levné knihy

Tento odkaz lze vyhledávačem interpretovat tak, že uživatel vytvářející tento odkaz je přesvědčen, že stránka na doméně andy.cz je relevantní ke klíčovému spojení „levné knihy“. Vyhledávače tedy analyzují texty odkazů, aby mohli vypočítat, jaké slova si uživatelé s danou stránkou spojují. Jestliže spousta významných stránek uvádí například, že na stránkách andy.cz jsou levné knihy, tato stránka je pak velmi relevantní pro případný vyhledávací dotaz „levné knihy“ a vyhledávače tedy zohlední tento fakt při tvorbě SERP. Toto hledisko je však ovlivnitelné a, jak bude ukázáno později, často využívané právě SEO spamery. Proto se snaží vyhledávače v rámci posuzování sémantiky jednotlivých odkazů zahrnout i tzv. okolí odkazu. Jedná se o další informace, které je vyhledávač schopen ze stránky získat a které mu pomohou rozhodnou, jak moc je informace uvedená v anchoru odkazu důvěryhodná. Tyto informace vyhledávač získává většinou buď z analýzy bezprostředního okolí odkazu, z analýzy celkového obsahu odkazující stránky nebo z analýzy provázanosti odkazující stránky mezi dalšími stránkami.

2.2.2.3 Analýza bezprostředního okolí odkazu

Jedná se o nejrychlejší a nejjednodušší způsob určování důvěryhodnosti anchoru odkazu. Vychází z logiky, že odkaz je na stránce zasazen do určitého kontextu, a tato metoda se snaží tento kontext analyzovat. V nejjednodušší verzi může tato metoda analyzovat, jestli se vůbec odkaz nachází v nějakém kontextu (například ve větě) anebo je pouze umístěn do hnízda dalších odkazů. V tomto případě lze usuzovat, že informace obsažená v odkazu zasazeném do kontextu je důvěryhodnější než jen samotně vystavený odkaz (který může indikovat snahu o SEO spam). Pokročilejší metody analyzující okolí odkazu se snaží určovat i relevantnost okolí k danému anchoru a další charakteristiky. [3]

2.2.2.4 Analýza celkového obsahu odkazující stránky

Tato metoda je podobná předchozí analýze bezprostředního okolí odkazu, ale s tím rozdílem, že vyhledávač se v tomto případě snaží rozpoznat tématikou příbuznost odkazující stránky vzhledem k textu v anchoru či k samotné odkazované stránce. Používá se zde nejčastěji již vypočítaných hodnot obsahových kvalit dané stránky uvedených v článku 2.2.1, v pokročilejších vyhledávačích se může do výpočtu zahrnout celkové hodnocení obsahu stránky získané nejen z hodnocení obsahu (on-page faktorů) ale i z hodnocení zpětných odkazů (off-page faktorů) samotné odkazující stránky. [3]

2.3 Search Engine Optimalization

Sear Engine Optimalization (SEO) je souhrnný název pro činnost, jejímž cílem je zlepšit pozice určité stránky ve vyhledávačích za účelem získání většího počtu návštěvníků (tzv. traffic). Jedná se o mnoho navazujících činností, které zahrnují i plánování a vzájemně se mohou proplétat. Z marketingového hlediska je SEO podmnožinou SEM (Search Engine Marketing) a rozumí se tímto investice do optimalizací stránky a do navazujících činností (například nákup zpětných odkazů). V širším slova smyslu se většinou hovoří SEO uvažované stránky jako o souhrnu tzv. on-page a off-page faktorů. Tyto faktory lze více či méně ovlivňovat právě pomocí metod SEO a do značné míry určují, jestli se daná stránka zobrazí v SERP na požadovaný dotaz či nikoliv.

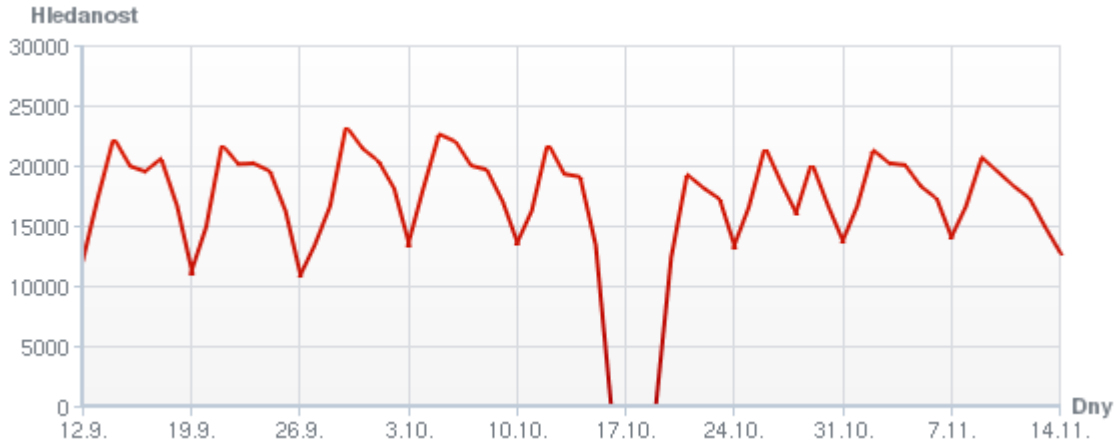
2.3.1 SEO z pohledu internetového marketingu

SEO je v dnešní době spojované především právě jako možná investice do internetové reklamy. Podle výzkumů efektivity různých reklamních možností internetu je totiž SEO nejzajímavější internetovou reklamou z hlediska počtu relevantní přivedených zákazníků.[4] Tato efektivita se sleduje většinou jako poměr cena reklamy / počet cílových konverzí u uživatelů z daného reklamního zdroje (pro tyto účely lze například použít volně dostupný nástroj Google Analytics).

Investice do SEO je výhodná oproti jiným reklamním systémům (například banerové reklamy) především proto, že optimalizovaná stránka se zobrazí uživatelům, kteří mají úmysl navštívit stránku ze SERP a do značné míry vyhledávači důvěřují, že ve výsledcích vyhledávání je právě stránka, kterou hledali. Takto získaný uživatel má tedy všeobecně větší konverzní poměr, jelikož aktivně hledal požadovanou informaci a důvěřuje svému vyhledávači. Uvažovaný uživatel totiž předpokládá, že SERP je v co největší míře objektivní a není ovlivněn komerčními faktory. Určitou podobnost lze vidět na PPC reklamních systémech, které sice mohou reagovat na vyhledávací dotaz uživatele a být tedy relevantní, nicméně míra relevance je značně sražena faktem, že pozice i texty v těchto systémech jsou ovlivněny komerčními aspekty (lepší pozici získává většinou lépe zaplacená reklamní sdělení).

Uvažujme modelový případ Search Engine Marketingu (SEM) například realitní společnosti z Brna, jejíž internetová prezentace není viditelná na pro tuto společnost zajímavá klíčová slova, a proto zvažuje investici do SEO. Pravděpodobně nejzajímavější by bylo, kdyby se této společnosti podařilo získat přední pozice na často vyhledávaná a relevantní klíčová slovní spojení (např. reality, reality Brno, prodej bytů Brno, atd.) Objem vyhledávání těchto slov lze snadno zjistit například pomocí volně dostupných statistik vyhledávače seznam.cz (viz obrázek 2-1):

Rozšířená shoda (dotazy obsahující "reality") [5]



Obrázek 2-1 statistika dotazu z vyhledávače seznam.cz

	Za den	Za týden
Minimální	0	67 042
Maximální	23 170	132 770
Průměrná	16 868	118 075

Nejhledanější dotazy obsahující "reality" Od 13. 09. 2009 do 14. 11. 2009

	Hledaný výraz	Rozšířená shoda	Přesná shoda
0	reality	16 868	1 372
1	cz reality	3 367	1 422
2	reality www	1 139	17
3	cz reality www	918	506
4	reality remax	569	114
5	mm reality	551	495
6	cz reality remax	426	379
7	m reality	397	32
8	reality s	322	144
9	brno reality	292	216
10	reality sting	248	228

Z těchto statistik vyplívá, že některá klíčová spojení jsou pro uvažovanou společnost velmi zajímavá a vzhledem k objemu vyhledávání lze říci, že mohou být nosným pilířem celého internetového marketingu (v tomto případě by se pravděpodobně jednalo o relevantní a zároveň vyhledávané spojení „reality Brno“. Na tyto slovní spojení uvažovaná společnost již provozuje PPC

reklamní sdělení, které ji stojí měsíčně **e** (expenses) Kč a přivede $\boldsymbol{u}$ (unique users) zákazníků, jejichž průměrný konverzní poměr je $\boldsymbol{c}$ a průměrný zisk na jedné konverzi je $\boldsymbol{p}$ (profit). Míru efektivity této reklamy tzv. ROI (Return Of Investment) lze vypočítat pomocí následujícího vzorce:

$$\boldsymbol{ROI = u.c.p\ /\ e}$$

Marketingové oddělení této společnosti by tedy pravděpodobně utrácelo určité množství peněz za SEO a sledovalo by pomocí analytických nástrojů a tohoto vzorce, jak se jim tato investice vyplácí.

2.3.2 Etické metody SEO

Existuje spousta způsobů, které se používají pro zviditelnění stránek ve výsledcích vyhledávání. Tyto metody používají jak specializované společnosti zabývající se SEO/SEM tak tvůrci stránek či (občas i nevědomky) běžní uživatelé. Tyto metody lze rozdělit na tzv. etické a neetické (zakázané).

Etické metody jsou ty, které se snaží stránku snadněji zpřístupnit robotům vyhledávačů a zároveň usnadňují samotné použití stránek pro uživatele. Jedná se především o změny on-page faktorů za cílem správného rozlišení obsahových hodnot stránky.

Tyto metody jsou většinou přímo doporučovány samotnými vyhledávači, neboť jejich použití usnadňuje vyhledávačům určení „čím se stránka zabývá“. Proto některé vyhledávače umožňují webmasterům analyzovat on-page faktory jejich stránek (například nástroj google webmaster tools). Vyhledávač seznam.cz v současné době nedisponuje obdobným nástrojem, nicméně obsahuje nápovědu „Jak lépe optimalizovat“. Jedná se o všeobecný soupis praktik, které by webdesigneři měli dodržovat při tvorbě webu. Tyto praktiky se označují jako etické metody SEO a jsou uznávány googlem i dalšími vyhledávači. Lze je rozdělit na on-page a off-page optimalizace

2.3.2.1 On-page etické optimalizace

Mezi on-page faktory lze řadit veškeré informace, které stránky obsahují (nejen viditelné, ale i části zdrojového kódu). V rámci SEO optimalizace se autoři stránek snaží většinou o tyto úpravy stránek:

- Přizpůsobení titulku stránky optimalizovanému výrazu
- Použití požadovaného výrazu v H1
- Zvýraznění tohoto výrazu ve všech výskytech v textu (pomocí)
- Přizpůsobení linkové struktury webu pro zvýšení počtu odkazů na optimalizovanou stránku (snahou je, aby optimalizovaná stránka byla s webem v co největší míře propojena, tedy každá stránka by měla odkazovat na tu optimalizovanou)
- Změna anchorů těchto zpětných odkazů (například místo „úvodní stránka“ je vhodné použít anchor který obsahuje přímo optimalizované klíčové slovo)

- Použití optimalizovaného slova v alternativních popisech obrázků (například v případě proniknutelného loga společnosti směřující na úvodní stránku)

2.3.2.2 Off-page etické optimalizace

Z hlediska toho, jak moc je etické ovlivňování off-page faktorů se názory uživatelů i postoje vyhledávačů různí. Zatímco google nedoporučuje, aby byly uměle ovlivňovány off-page faktory stránek (například pomocí budování zpětných odkazů – tzv. linkbuilding). [6] Seznam.cz se k výměně zpětných odkazů staví spíše neutrálně, dokonce doporučuje, čeho se při výměně odkazů vyvarovat. [7]

V dnešní době se často praktikuje právě výměna zpětných odkazů. Princip této metody spočívá v tom, že na sebe vzájemně odkazují 2 weby, ideálně s podobnou tématikou. Všeobecně platí, že web by měl odkazovat jen na relevantní a důvěryhodné cizí weby a pro výměnu odkazů by se neměl používat žádný automatický systém. Smysl odkazů v tomto případě tvoří určité doporučení autora webové stránky o kvalitě jiné webové stránky. Lze tedy takto propojit své partnerské weby či poukázat na zdroj, že kterého jsme čerpali.

Výměna zpětných odkazů by neměla být samoúčelná a v rozumné míře. Znamená to, že snaha o zmanipulování vyhledávače pomocí velkého množství odkazů je považována za neetickou a většinou i tvrdě postihována.

Další oblíbenou a poměrně účinnou metodou jak získat zpětné odkazy je tzv. zápis do katalogů. Jedná se o placené či neplacené zápisy v internetových katalozích, jejichž smysl je vytvořit seznam stránek roztříděných podle různých kritérií (například podle oborů, velikostí firem, atd.). Čím více katalogů obsahuje optimalizovanou webovou stránku, tím více daný web de facto získá zpětných odkazů. Zde je však nutné dávat pozor, na který katalog se firma zapisuje. Některé katalogy totiž neplní jinou funkci než poskytování odkazů a nepřináší žádné kvalitativní hodnoty – tyto katalogy v podstatě automaticky přijímají jakýkoliv zápis bez ruční kontroly administrátorem a informace v těchto katalozích mohou být značně nedůvěryhodné a účelné. Lze předpokládat, že vyhledávače zpětné odkazy z takovýchto katalogů ignorují nebo dokonce postihují. Existují programy, které dokážou provádět tyto katalogové zápisy automaticky do stovek katalogů (například mezi seařskou komunitou velmi oblíbený program SubmitWolf). Použití takovýchto programů je však při odhalení vyhledávači trestáno.[8] Všeobecně je tedy doporučeno provádět zápisy pouze ručně a to jen do kvalitních a ověřených katalogů.

2.3.3 Zakázané metody SEO

Mezi zakázané (nebo též neetické) metody SEO se řadí všechny metody, jejichž použití hodnotí vyhledávače negativně (většinou postihem, který znamená snížení pozic v rámci SERPu či tzv ban – vyřazení stránky z výsledků vyhledávání). Tyto metody lze rovněž rozdělit na on-page a off-page.

2.3.3.1 Zakázané on-page optimalizace

Snahou těchto optimalizací je většinou uměle měnit obsahové faktory stránek. Jedná se především o tyto metody:

- Skrytý text - nejčastějším případem je schovávání uměle přidaných slov, jejichž smyslem je zvýšení density a relevance stránky na dané slovo. Skrývání se provádí nejčastěji pomocí kaskádových stylů (například nulová výška fontu) nebo barva textu splývá s pozadím. V současné době jsou vyhledávače schopny skrývání textu pomocí CSS rozpoznat, horší situace je pokud barva textu je velmi podobná barvě obrázku v pozadí. Odhalení skrytého textu některé vyhledávače (např. google) trestají vyřazením z indexu. [9]
- Uměle přidaný SEO text - tato metoda se rovněž snaží uměle zvýšit relevanci dané stránky na určitý výraz. Většina vyhledávačů hodnotí negativně podezřele velkou míru optimalizace stránky na jedno slovo (například příliš vysoká densita slova). Další nevýhodou této metody je fakt, že stránky jsou hůře čitelné pro samotné uživatele a tedy ve výsledku méně kvalitní. Právě z tohoto důvodu se snaží vyhledávače příliš optimalizované stránky rozpoznávat a trestat.
- Skryté odkazy – smysl skrytých odkazů je většinou přivést robota vyhledávače na stránky, které jsou určeny pro zvýšení obsahové hodnoty stránek nebo jako zásobárna zpětných odkazů (viz uměle vygenerovaný obsah) nebo odstranit esteticky negativní efekt odkazů na jiné stránky (tzv. SEO lišty)
- Nepřímé přesměrování a maskování – jedná se o metodu, kdy je uživatelům předkládán jiný obsah než vyhledávači. Může se jednat například o zobrazení textu HTML vyhledávačům, zatímco uživatelům se zobrazuje objekt flash s rozdílným obsahem, maskování pomocí obrázků nebo přesměrování, kdy je uživatel přesměrován na stránku s jiným obsahem pomocí javascriptu.
- Doorway-pages – neboli tzv. portálové stránky jsou soubory stránek, kde je každá stránka vysoce optimalizovaná právě na jeden konkrétní výraz. Tímto se portálové stránky jsou schopny dostat do SERPu na mnoho klíčových výrazů a následně pomocí podvodného přesměrování přesměrovat uživatele na jinou stránky, které doorway-pages zvyšují traffic.
- Cloaking – jedná se o metodu, jejímž smyslem je rovněž vyhledávačům poskytnout jiné informace než samotným uživatelům. Tato metoda využívá nejčastěji službu, která běží na serveru hostujícím dané stránky. Tato služba je schopná pomocí různých parametrů (například IP adresa, rychlost průchodu stránkami atd) zjistit, jestli stránky navštívil reálný uživatel nebo robot vyhledávače a podle toho odpovídat různými stránkami. Smyslem této metody je nejčastěji podstrčit vyhledávači mnoho (až tisíce) nekvalitních často uměle generovaných nicméně vysoce optimalizovaných stránek. Pro rozpoznání této metody

vyhledávače používají 2 průchodů stránek – nejprve je stránka navštívena standardním robotem vyhledávače, poté vyhledávač simuluje návštěvu uživatele. Následně jsou porovnány získané obsahy, pokud se liší, tak se s největší pravděpodobností jedná o cloaking a stránka bývá v drtivé většině zcela vyřazena z indexu vyhledávače.

- Duplicitní obsah – duplicitním obsahem vyhledávače rozumí části textů webu, které se v rámci jedné či několika domén opakují. Tento jev nemusí být ještě nutně podvodnou aktivitou, jelikož části stránek se mohou přirozeně opakovat (například v diskuzních fórech, různých verzích webových stránek, atd.). Nicméně se lze poměrně často setkat s případy, kdy obsah webových stránek je v různých modifikacích duplikován napříč doménami, aby se zvýšila pravděpodobnost, že uživatel navštíví stránku s naším obsahem. Tato metoda může vést k tomu, že vyhledávač zobrazí několik výsledků hledání, ve kterých je však téměř stejný obsah. Toto snižuje kvalitu vyhledávání, a proto se vyhledávače snaží duplicitní web vyřadit ze svého indexu (respektive ponechat pouze jednu instanci daného webu – pro tyto případy používá google tzv. kanonizaci [10]. Na internetu se lze rovněž setkat s modifikací této metody, kdy stránky duplikují cizí obsah, aby získali co nejvíce relevantního kvalitního obsahu. Použití této metody je nejen opět postihováno vyhledávači, ale rovněž považováno za porušení Zákona o duševním vlastnictví.
- Uměle vygenerovaný obsah – jedná se stránky, které nejsou nijak přínosné pro uživatele a jejichž existence je motivovaná pouze cílem ovlivnit vyhledávače. Tyto stránky většinou obsahují náhodné odstavce textů přebraných z jiných stránek nebo vytvořených pomocí slovníků. Tento text nedává smysl, nicméně obsahuje v dostatečné míře požadovaná klíčová slova. Smyslem těchto stránek není získat pozice v SERPu, ale poskytovat dostatečnou základnu (většinou tisíce stránek) tematických zpětných odkazů. Pokročilou modifikací této metody je použití automatických překladačů webových stránek pro překlad stránek napsaných v cizím jazyce do požadovaného jazyku (například překladač yahoo nebo google). Takto přeložené stránky jsou následně ukládány na nové domény či subdomény a používány obdobně jako vygenerované stránky. V rámci této práce jsem implementoval experimentální generátor stránek s umělým obsahem. Tento program bude použit pro potřeby testování SEO spamu.

2.3.3.2 Zakázané off-page optimalizace

V současné době vyhledávače považují při sestavování SERPu off-page faktory jednotlivých stránek za jedno z rozhodujících hledisek. Důvodem tohoto hodnocení je fakt, že off-page faktory se nedají tak snadno ovlivnit jako on-page faktory a mají přirozenou vypovídací hodnotu o kvalitě stránky

(množství zpětných odkazů de facto uvádí množství tvůrců stránek, kteří jsou přesvědčeni, že odkazovaná stránka je pro ostatní uživatele rovněž zajímavá).

Tohoto faktu jsou si samozřejmě rovněž vědomi samotní optimalizátoři či společnosti zabývající se SEO marketingem. Proto byly vyvinuty následující metody, jejichž snahou je tzv. linkový spam – tedy nepřirozeného získání zpětných odkazů ve snaze zmanipulovat hodnocení vyhledávačů:

- Tapetování katalogů – jedná se o metodu, která využívá faktu, že zápisy do některých katalogů nejsou kontrolovány administrátory, a že některé katalogy dovolují zapsat více stránek z jedné domény. Takto lze do katalogu zaregistrovat jednotlivé podstránky daného webu a získat tím velké množství zpětných odkazů. Dle mých experimentů v českém internetu existuje cca 25 katalogů, které lze zneužít pro tapetování. Vyhledávače se většinou proti této metodě brání vyřazením katalogů, které lze tapetovat, z indexu.
- Linkfarmy – jedná se o propojení webů, které na sebe vzájemně odkazují. Tyto weby lze vytvářet ručně, ale většinou se jedná o automatické systémy, které umožňují automatické prolinkování všech zúčastněných webů. Tyto linkové farmy umožňují stránkám v rámci farmy získat velké množství odkazů a zároveň může sloužit jako zásobárna odkazů pro weby mimo farmu. Existují i další modifikace linkfarem, které obsahují provizní systémy atd. (nejznámějším českým případem je web www.lista.cz). Členství v linkfarmách většinou vyhledávače postihují penalizací.
- Spamování blogů a fór – tato metoda je založená na použití webů, které umožňují vkládání odkazů do svých příspěvkových systémů, fór atd. Pro tuto metodu se používají většinou automatické systémy, které prochází fóra a zkouší jejich zranitelnost vůči automatickému vkládání příspěvků s odkazy (například program Instant Blog Submitter). Proti této metodě by měli být zabezpečeny především samotné blogy a fóra, pokud vyhledávače zaznamenají podezřele vysoký výskyt stejných odkazů z jednoho fóra, může být toto fórum penalizováno a váha těchto odkazů nulována.
- Scraper site – jedná se o stránky, které používají obsahy jiných stránek a slouží často jen jako zásobárna zpětných odkazů pro další stránky. Tyto stránky mohou rovněž být optimalizovány, aby se zvýšila tématická hodnota těchto zpětných odkazů. V některých případech obsahují stránky PPC reklamní systémy třetích stran – tyto stránky se často označují jako MFA (Made For AdSence).
- Pronajímání zpětných odkazů – síla zpětného odkazu je odvozena od důležitosti webu, ze kterého zpětný odkaz vede (většinou se zde hodnotí jeho rank, případně jeho tematičnost, atd.) Majitelé vysoce hodnocených webů mohou považovat za velmi zajímavý způsob dalšího zisku tzv. pronajímání zpětných odkazů. Ceny odkazů se pohybují od 30 do 3.000,- Kč / měsíc podle následujících faktorů: rank stránky, počet zaindexovaných stránek ve vyhledávači (lze zjistit např. dotazem site:), délka dohodnutého pronájmu

(většinou 1-6 měsíců) a počet externích odkazů na dané stránce. Tyto odkazy jsou většinou textové a jejich anchor obsahuje optimalizované slovo. Odkazy jsou nejčastěji tzv. site-wide – to znamená, že odkaz je umístěn na všech podstráních dané stránky (proto se při určování ceny zpětného odkazu často zohledňuje, kolik stránek daný web obsahuje). Pronajímání zpětných odkazů je v současné době nejpopulárnější způsob získávání zpětných odkazů. Touto metodou lze získat tisíce zpětných odkazů z vysoce kvalitních webů – což do značné míry ovlivní hodnocení daného webu vyhledávači. Dalo by se říci, že pronájem zpětných odkazů je nové business odvětví, které využívá internet. V rámci českého internetu existují 2 systémy umožňující pronájem zpětných odkazů: www.mlink.cz a www.directoglobal.com. Uživatelé internetu se však domlouvají na pronájmech prakticky všemi dostupnými informačními kanály, proto v podstatě nelze odhalit, které odkazy jsou přirozené a které pronajaté. Jelikož pronajaté odkazy do značné míry negativně ovlivňují pořadí stránek v SERPu a tedy i kvalitu vyhledávání, vyhledávače se snaží proti tomuto efektu bojovat. Například vyhledávač google zaznamenává rychlost nárůstu zpětných odkazů, pokud množství zpětných odkazů roste podezřele rychle, je tento web penalizován. [11] Tento vyhledávač rovněž vybízí uživatele, aby hlásili veškeré podezření na cizí pronajaté zpětné odkazy či případné své pronajaté odkazy pomocí nástroje google webmaster tootls. [12] Problém pronajatých odkazů je v současné době největší výzvou pro vyhledávače a proto se i tato práce snaží navrhnout metodu, která zmírní dopad pronajatých odkazů.

2.3.4 Definice SEO spamu

Na základě předchozích poznatků můžeme definovat SEO spam jako soubor neetických on-page a off-page optimalizací, jejímž smyslem je zmanipulovat vyhledávače pro získání lepších pozic optimalizované stránky ve výsledcích vyhledávání SERP. Tyto metody všeobecně negativně ovlivňují kvalitu vyhledávání a je tedy žádoucí, aby vyhledávače použití těchto metod detekovali a trestali.

3 Metody obrany proti SEO spamu

Problémem obrany proti SEO spamu je fakt, že spameři neustále vyvíjejí a zdokonalují své spamovací metody. Na internetu jsou zdarma dostupné tisíce nástrojů, které jsou schopny analyzovat nejrůznější faktory stránek, automatizovat výměnu zpětných odkazů a provádět další činnosti, které profesionální společnosti nabízející SEO/SEM používají každý den.

Dalo by se říci, že vývojáři vyhledávačů jsou ve výhodě, jelikož spameři nikdy nemohou dopředu vědět (za předpokladu, že nedojde k úniku informací ze strany zaměstnanců daného

vyhledávače), jaké změny a metody proti SEO spamu se budou implementovat. Tato výhoda je však velmi relativní, jelikož spameři mají jednoduše drtivou početní převahu a jsou tedy schopni velmi rychle na případné změny reagovat (například pomocí reverzního inženýrství).

Všeobecně lze tedy říci, že vyhledávače své metody obrany proti SEO spamu do značné míry tají, proto tato práce poukazuje na veřejně dostupné metody.

3.1 Všeobecné požadavky na kvalitu vyhledávání

Vyhledávání v síti internet sebou nese mnoho úskalí. Jedná se především o fakt, že nelze s určitostí měřit kvalitu výsledků vyhledávání, jelikož relevantnost určité stránky k danému vyhledávanému výrazu je velmi subjektivní. Názory jednotlivých uživatelů internetu mohou být též protichůdné a i stránka, která zneužívá SEO spam, může být pro samotné uživatele přínosná (zde se především setkáváme s pravidlem, že například eshopy, které si mohou dovolit investovat desetitisíce korun měsíčně do linkového spamu, jsou většinou dostatečně velké, aby byly pro uživatele internetu zajímavé). Vyhledávače musí být dostatečně neutrální, aby jejich výsledky vyhledávání byly přínosné pro co největší komunitu lidí. Dalším omezujícím parametrem moderních vyhledávačů je rychlost odezvy, ta by se měla pohybovat v řádu milisekund, což do značné míry omezuje pokročilé algoritmické možnosti (například personifikované hledání).

3.1.1 Problém určení autoritativních domén

Počátkem každého vyhledávání je formulace uživatelova dotazu. Ten může mýt 3 podoby:

- Specifický dotaz (například „umožňuje SBS 2003 terminálový server?“)
- Všeobecný dotaz (například „programovací jazyk Python“)
- Dotaz na podobnou stránku (například „najdi stránku podobnou python.org“)

Každý uvedený dotaz vyžaduje jiný přístup. Zatímco problém prvního příkladu je fakt, že existuje pouze velmi málo stránek, které opravdu obsahují požadovanou informaci (pokud vůbec), na všeobecný dotaz nalezne vyhledávač pravděpodobně příliš mnoho stránek obsahující požadovanou informaci a musí tedy z těchto stránek vybrat ty nejlepší. Dalším problémem je fakt, že mnohé stránky, které jsou velmi relevantní na daný hledaný výraz (uvažujme například výraz „výrobce automobilů ČR“) tento výraz ani nemusí na svých stránkách obsahovat. Nelze očekávat, že například automobilka Škoda bude mít na svých stránkách tuto informaci. Vyhledávače tedy nemohou spoléhat pouze na analýzu sémantiky stránek, ale musí být schopny rozpoznat i jejich důležitost a věrohodnost. – tedy určit tzv. míru autoritativnosti domény.

3.1.1.1 Pagerank algoritmus

PageRank je pravděpodobně nejznámější algoritmus, ze kterého vychází moderní vyhledávače (např. Google). Tento algoritmus vychází z algoritmu pro sčítání citací v publikacích a byl navržen roku 1998 dvěma autory – Brian a Page (právě od tohoto autora pochází název PageRank).

Principem tohoto algoritmu je výpočet číselného skóre pro každé stránky pomocí analýzy hypertextových odkazů směřujících na danou stránku. Algoritmus počítá číselné váhy každého hypertextového odkazu za účelem ohodnocení jeho důležitosti, část této váhy stránka opět předává stránkám, na které odkazuje, a tak dále. U každé stránky můžeme zjistit tzv. Google Toolbar PageRank. Jedná se o hodnotu z intervalu 0-10, která je odvozena z logaritmického přepočtu skutečného PageRanku. Tato hodnota je přepočítávána cca 5x do roku [13].

Předpokládejme stránku A, na kterou odkazují stránky T_1 ... T_n. Samotný výpočet PageRanku pro tuto stránku PR(A) lze zapsat pomocí následujícího vzorce [14]:

$$PR(A) = (1-d)/m + d * (PR(T_1)/C(T_1) + \ldots + PR(T_n)/C(T_n))$$

d – damping factor (faktor útlumu) $d \in (0,1)$, většinou se uvažuje $d = 0{,}85$ [15]

m - celkový počet zaindexovaných stránek.

$C(T)$ – počet odkazů, které vychází ze stránky T

Tento výpočet se provádí v iteracích, přičemž za $PR(T_i)$ se berou hodnoty PageRanku stránek z minulé iterace výpočtu. Na začátku iterace se mohou zvolit libovolné vstupní hodnoty, jelikož tento vzorec po několika iteracích konverguje.

V původním návrhu vyhledávače Google byly stránky řazeny do SERPu podle velikosti jejich PR [15]. V dnešní době již PageRank není tak důležitý a jedná se pouze o informativní hledisko, jejíž kvantitativní vliv na pořadí stránky v SERPu je sporný [16].

3.1.1.2 HITS algoritmus

Tento algoritmus byl poprvé popsán Jonem Kleinbergem v roce 1998. Myšlenka algoritmu HITS (Hyperlink Induced Topic Destilace) je založena na faktu, že autority a huby (stránky odkazující na autority) se vzájemně posilují. Váha autoritativní stránky je vypočtena jako součet váhy stránek hubů, které na ni odkazují. Zároveň váha hubu je vypočtena jako součet váhy autoritativních stránek odkazujících na tento hub.

Zápis algoritmu HITS je následující. Nechť S je množina stránek, pro které chceme vypočítát váhy hubu a autoritativních stránek, nechť n je počet stránek v této množině. Pak H je podmnožinou S obsahující stránky v roli hubu, a A je podmnožinou S obsahující autoritativní stránky. Protože každá stránka může být v roli hubu i autoritativní stránky, množiny A a H se překrývají. Pro každou stránku i v roli hubu je $F(i)$ počet odchozích odkazů. $B(i)$ je naopak počet příchozích odkazů pro

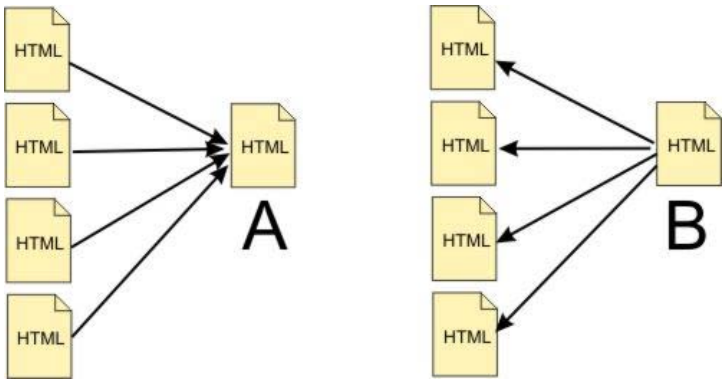
každou autoritativní stránku i. N-rozměrný vektor váhy autoritativní stránky je označován jako a zatímco vektor váhy hubu je označován jako h. Tyto váhy lze vypočítat podle následujících vzorců:

$$a_i = \sum_{j \in B(i)} h_j$$

$$h_i = \sum_{j \in F(i)} a_i$$

Tento proces je iterativní. Nejprve všechny váhy dostávají hodnotu 1, pak se postupně počítají další váhy stránek hubu a autorit. Tyto hodnoty jsou následně normalizovány. Tato fáze se opakuje, dokud vektory a a h konvergují.

Tento algoritmus má však řadu nedostatků. Například díky principu vzájemného posilování může vzniknout následující situace: uvažujme hub, který ukazuje na řadu autoritativních stránek (viz hub B na obrázku níže), a řadu hubů, které odkazují na jednu autoritativní stránku (autoritativní stránka A na obrázku). Pokud je počet autoritativních stránek, na nichž odkazuje B, větší než je počet hubů, pak algoritmus HITS přidělí všechny váhy na autoritativní stránky v pravé části obrázku a autoritativní stránka A dostane váhu téměř nulovou.



Obrázek 3-1 odkazování hubů na autoritativní stránky

Důvod tohoto nestandardního chování je fakt, že hub B bude zpočátku dostávat velmi vysoké skóre a propagovat jej mezi autoritativní stránky. Zároveň však huby na levé straně budou mít velmi nízké skóre, a tím pádem budou mít příliš nízké váhy – autoritativní stránka A bude tedy mít poměrně malé skóre i přes to, že si zaslouží víc.

3.1.1.3 Topic sensitive Page Rank

Tento algoritmu je vylepšením tradičního Page Ranku, který vychází z faktu, že míra přenosu váhy zpětných odkazů mezi dvěma stránkami by měla být ovlivněna mírou tematické příbuznosti těchto dvou stránek. Předpokládá se, že například stránky z bankovního sektoru budou opět odkazovat na podobné stránky (které se zabývají například konkrétními finančními službami) nikoliv však na oborově velmi odlišné stránky (řekněme například o dietách), a pokud takováto stránka odkazuje na takto oborově nesourodou stránku, jedná se s největší pravděpodobností o zaplacenou reklamu,

která by se neměla počítat do celkového hodnocení odkazované stránky. Výše zmíněný algoritmus HITS do značné míry také počítá s oborově příbuznými stránkami (huby a autority jsou oborově příbuzné), nicméně tento algoritmus nemůže počítat s příliš velkými množinami a omezuje se pouze na nejbližší sousedy (další odkazované stránky) – jinak by výpočet trval díky použité rekurzi neúměrně dlouho.

Topic sensitive Page Rank rozšiřuje klasický algoritmus Page Ranku o výpočet obsahového ovlivnění Page Ranku (PR) v době generování výsledků vyhledávání. Algoritmus nepoužívá jeden universální vektor hodnot PR, ale počítá několik vektorů PR pro každý tematický celek. Vytvoření vektorů PR pro každou možnou množinu témat by ale vyžadovalo rozsáhlé zdroje, proto tento algoritmus v praxi používá pouze 16 vektorů reprezentující nejvyšší kategorie projektu ODP (Open Directory Project – viz dmoz.org). Teoreticky mohou být i jiné zdroje témat, ale původně byl algoritmus navržen na těchto 16 vektorů, jelikož ODP je největší nekomerční projekt vytvořený a editovaný velkým počtem nezávislých dobrovolníků. Je tedy nepravděpodobné, že zařazení stránek do jednotlivých kategorií by byly ovlivněny 3. stranou. Pro každou stránku na webu algoritmus vypočítává skóre (Topic sensitive PR) ve vztahu k těmto tématům. Toto skóre je vypočítáno každé stránce dopředu (podobně jako klasický PR) a uloženo v režimu offline. V času konkrétního dotazu je určena příbuznost dotazovaného slova k dané tématice a konečný SERP je vytvořen kombinací Topic sensitive PR (vzhledem k určené tématice) s jinými metrikami (např. analýzy obsahu).

Samotný algoritmu lze popsat následovně: nechť c_j je jedna z 16ti kategorií dle ODP; pro každé téma c_j je nezbytné vypočítat PageRank vektor. Nechť M je upravená matice sousednosti. Každý prvek m_{ji} má hodnotu $1/N_j$, existuje-li odkaz ze stránky j do stránky i a kde N_j je počet odchozích odkazů na stránce j. V případě, že neexistuje zpětný odkaz z j do i je hodnota elementu v této matici rovna 0. Pak původní PageRank vzorce v notaci s uvažovanou maticí vypadá takto:

$$PR = (1-\alpha)(D+M)PR + \alpha p$$

Kde: $p = \left[\frac{1}{n}\right]_{n\times 1}$ - uniformní vektor útlumu

$D = p \times d^T$ - d^T je transponovaný n-dimendionální vektor sloupců, který identifikuje stránky neobsahující odchozí odkazy (tzv. zero-out degree)

Poměrně malou úpravou tohoto vzorce lze počítat tematický PageRank. Nechť T_j je množina stránek souvztažná k tématu c_j. Pak můžeme použít místo uniformního vektoru útlumu p (z výše uvedeného vzorce) neuniformní vektor v_j:

$$v_{ji} = \begin{cases} \frac{1}{|T_j|} & i \in T_j \\ 0 & i \cap T_j = \emptyset \end{cases}$$

Topic sensitive PageRank navíc počítá ještě v rámci každého tématu c_j vektor D_j. Jednotlivé prvky D_{jt} tohoto vektoru obsahují počet výskytů každého termínu souvisejícího s tématem c_j. Pro určení tématu, ke kterému se vyhledávací dotaz vztahuje, je třeba uvažovat tyto 2 případy:

- V prvním případě uživatel zdůrazňuje klíčové slovo na stránce a začne hledat. V tomto případě je téma definováno obsahem této stránky. Například jestliže uživatel označil na stránce o slavných stavbách slovo architektura, ve výsledcích vyhledávání by se již neměli vyskytovat stránky zabývající se hardwarovou architekturou. Takže pokud termín q je zvýrazněn na uvažované stránce u, bude jeho kontext q' obsahovat slova v u.
- Ve druhém případě uživatel zadá klíčové slovo do vyhledávacího formuláře konvenčním způsobem. V tomto případě kontextem dotazu q je hledaný výraz sám o sobě: $q' = q$. Dalším přizpůsobením je udržovat historii hledaných termínů a tyto pak použít jako kontext q'. Blízkost kontextu dotazu q' k jednomu z témat c_j lze vypočítat během výpočtu odpovědi pomocí následujícího vzorce [20]:

$$P(c_j|q') = \frac{P(c_j) \cdot P(q'|c_j)}{P(q')} \propto P(c_j) \cdot \prod_i P(q'_i|c_j)$$

Přibližná hodnota $P(q' \mid cj)$ se počítá s použitím vektoru D_j. Skóre pro jednotlivé hodnoty jsou počítány pro každou stránku d z indexu, která obsahuje původní hledaný výraz q. Pro každý dokument d sečteme blízkost k danému tématu – $P(c_j|q')$ a PageRank r_d, který je elementem vektoru $PR(\alpha, v_{ji})$ vztahovanému k tématu, do kterého stránka d náleží. Finální seřazení stránek do SERPu je pak určeno podle velikosti s_d:

$$s_{qd} = \sum_j P(q' \mid cj) \,.\, r_{jd}$$

Vzhledem k faktu, že PageRank je předpočítán dopředu, je tento algoritmus schopen počítat tematické přizpůsobení vyhledávání v reálném čase.

V současné době se spekuluje, že Google používá mnohem více vektorů se stromovou hierarchií – toto vyplývá ze samotného rozdělení slov ve veřejně dostupných statistikách. Takto vypadá například rozdělení do kategorií pro vyhledávaný dotaz „motory“ [19]:

$$Kategorie: Automobilový\ průmysl\ (25–50\ \%), Hry\ (10–25\ \%), Internet\ (0–10\ \%),$$
$$Životní\ styl\ (0–10\ \%), Rekreace\ (0–10\ \%)$$

Je tedy pravděpodobné, že vyhledávač google uvažuje jednotlivé váhy kategorií podle hledaného výrazu.

3.2 Metody detekce SEO spamu

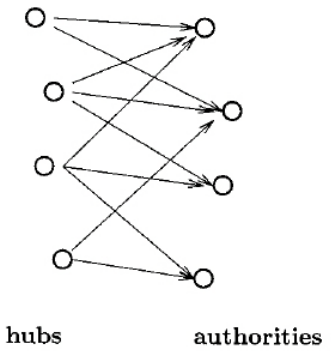
Spameři se mohou snažit ovlivnit vyhledávače buď tím, že uměle zvýší autoritativnost své domény, nebo úpravou on-page faktorů samotných stránek. Jelikož úprava on-page faktorů je poměrně jednoduchá a může ji provést „každý“, moderní vyhledávače již nekladou takový důraz při posuzování těchto faktorů. Proto je v současné době důraz vývoje metod detekce SEO spamu směřován na tzv. linkový spam, tedy umělé zvýšení autoritativnosti určité stránky.

3.2.1 Známé metody využívající analýzu hypertextových odkazů

3.2.1.1 Computing Hubs and Authorities

Tato metoda je založena na faktu, že lze najít množinu stránek relevantních vzhledem k hledanému výrazu (tzv. huby) a tyto stránky většinou odkazují na stránky, které jsou k tomuto výrazu autoritativní (autority). Existuje celá řada metod vycházející z tohoto principu, problémem je vždy nalezení dostatečného množství kvalitních hubů.

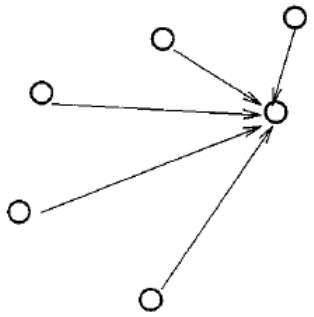
Nejčastěji se vychází z malého počtu stránek (například 200), které obsahují hledaný výraz a sleduje se na které další stránky tyto stránky odkazují – zde se předpokládá, že tyto odkazované stránky jsou rovněž relevantní huby. Dále lze rekurzivně sledovat, kam odkazují tyto huby, a vytvořit tak dostatečně velký graf stránek se závislostí hubů a autoritativních stránek (hubs – authorities) – tuto závislost demonstruje následující obrázek 3-2 [21]:



Obrázek 3-2 závislost mezi huby a autoritami

Předpokládá se, že kvalitní huby odkazují na autoritativní stránky (ty však většinou již na huby neodkazují). Kvalitní hub se pozná tím, že odkazuje na autoritativní stránku, zároveň kvalitní autoritativní stránka je indikována huby, které na ni odkazují. Lze tedy vidět, že huby a autoritativní stránky jsou v určité synergii. Z dostatečné analýzy nalezeného grafu stránek lze tedy vyhledat tuto synergii a rozpoznat jak kvalitní huby, tak kvalitní autoritativní stránky. Z těchto autoritativních stránek se pak tvoří podle množství jejich hubů výsledný SERP.

Tato metoda je poměrně úspěšná při boji proti linkovému spamu. Spameři mohou získat velké množství zpětných odkazů (například pronájmy z cizích serverů či budováním linkových farem), nicméně tyto stránky jsou většinou nerelevantní. Tuto situaci lze pro srovnání s modelem hubs-authorities znázornit následujícím obrázkem:



Obrázek 3-3 ukázka pravděpodobného spamu

Z výše zobrazeného grafu (obrázek 3-3) je evidentní, že stránky, které jsou podporovány linkovými spamy se jen velmi těžko (pokud si to „nezaslouží“) dostanou mezi autoritativní stránky. Takto lze poměrně účinně odfiltrovat stránky používající linkový spam, na druhou stranu má tato metoda několik nevýhod. Jedná se především o způsob určení počáteční množiny hubů, která by sama o sobě neměla být moc zatížená spamem. Dalším problémem je sestavení opravdu odpovídajícího SERPu – zde narážíme především na fakt, že zohledňovat pouze množství podporujících hubů nestačí.

3.2.1.2 Trust Rank

Jedná se o metodu, jejíž snahou je přímo ohodnotit důvěryhodnost webu a podíl spamu na jejím ranku. TrustRank byl navržen v roce 2004 autory: Zoltán Gyöngyi, Hector Garcia-Molina (Standfordská Univerzita) a Jan Pederson (Yahoo) [22].

Tato metoda vychází z následujících předpokladů:

1) Autoritativní weby neodkazují (nebo jen ve velmi malé míře) na spamové stránky
2) Pečlivost, se kterou autoři stránek vybírají, na které další stránky jejich stránka odkazuje je inverzní vůči množství odkazovaných stránek
3) Spamové stránky často odkazují na další spamové weby

TrustRank vychází z odkazové analýzy, kdy se šíří tzv. Trust (důvěra) od manuálně označených autoritativních stránek (tzv. Seeds) dále internetem. Využívá se zde útlumu – to znamená, že stránky, na které odkazuje přímo seed mají největší trust, stránky na které odkazují, až tyto stránky mají o něco menší a tak dále. Rozložení linkového spamu v odkazovém grafu by tedy mohlo vypadat následovně:

Seeds (0% spam) –> 1 click from seed (0,01% spam) –> 2 clicks from seed (1,5%) –> 3 click from seed (15%) -> … –> SPAM

Tato metoda dokáže přesně vyjádřit poměr spamu v hodnocení celkového ranku stránky, její úskalí je však velikost a obsah množiny Seeds. Tato množina musí být totiž dostatečně velká a zároveň kvalitní, aby po aplikaci TrustRank algoritmu zasáhla spravedlivě veškeré stránky.

3.2.1.3 Anti-Trust Rank

Tato metoda je stejně jako Trust Rank zaměřena na přímé rozpoznávání SEO spamu pomocí odkazové analýzy. Tato metoda rovněž používá mechanismu inverzního pageranku, nicméně vychází z faktu, že špatné (spamové) stránky jsou mnohem více propojeny než dobré (autoritativní stránky). Spameři totiž velmi často používají linkové farmy, kdy jedna stránka odkazuje bezprostředně na své sousedy. Na rozdíl od Trust Ranku tato metoda tedy jako své počáteční stránky (seed) používá manuálně označené spamové stránky.

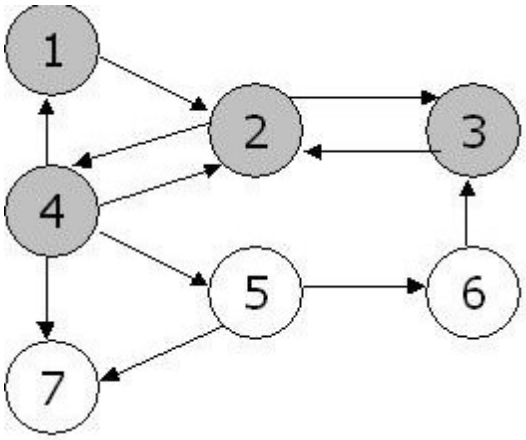
Stránky patřící do seedu jsou vybírány podle těchto 2 klíčových hledisek:

- Spamová stránka musí být velmi propojená s dalšími stránkami (nachází se na ní spousta externích odkazů na další stránky). Jelikož algoritmus Anti-Trust Ranku vychází z algoritmu inverzního Page Ranku, tak se tímto umožní, aby se dosáhlo velkého počtu stránek s pokud možno co nejmenším množstvím přeskoků (dalších kliků).
- Vybraná spamová stránka by měla mít sama o sobě pokud možno co největší Page Rank. Tento výběr je velmi užitečný, jelikož se tím postihnou další „důležité“ spamové stránky (dá se zde předpokládat, že tyto spamové stránky budou mít rovněž velký page rank -> vysokou pozici v SERPu).

Tato metoda rovněž vychází z principu, že případy, kdy dobrá (autoritativní) stránka odkazuje na spam jsou poměrně ojedinělé. Postup této metody je následující [23]:

1) Získání seed množiny stránek, které jsou ručně označeny jako spam a zároveň splňují výše uvedená kritéria
2) Výpočet transformované matice T webového grafu (odkazujících stránek)
3) Provedení algoritmu personifikovaného pageranku s použitím matice T na seed
4) Ohodnocení stránek s klesajícím skóre dle vzdálenosti od stránek označených jako seed nebo použití hodnoty treshold při jejíž překročení je stránka označena jako spam.

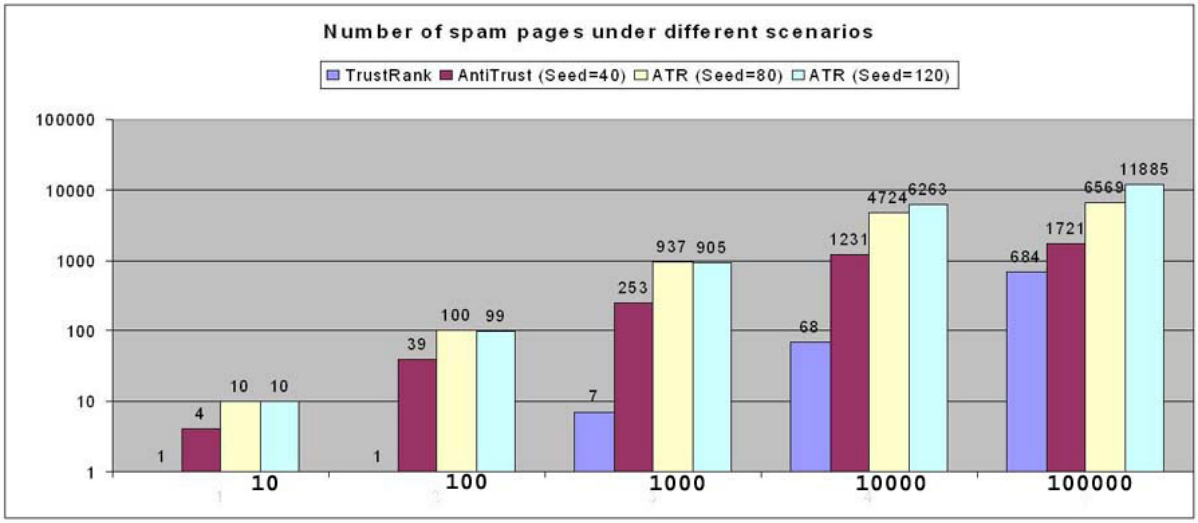
Mechanismus této metody lze demonstrovat na následujícím obrázku [23]:



Obrázek 3-4 šíření anti-trust ranku

Obrázek 3-3 znázorňuje jednoduchý graf odkazujících stránek, přičemž spamové stránky jsou vybarveny tmavě a kvalitní stránky jsou vybarveny bílou barvou. Předpokládejme, že jako seed byla označena stránka 1. Jelikož stránka 4 odkazuje na stránku 1, získá část z jejího Anti-Trust ranku. Ten se dále propaguje na stránku 2 a 3 s jistým útlumem oproti stránce 4 respektive 2. Jak lze očekávat, Anti-Trust rank by se měl propagovat i na stránku 6 (která již není spam), nicméně díky efektu útlumu tato stránka dostane již zanedbatelné Anti-Trust skóre. Metoda tedy správně označila spamové stránky.

Tato metoda je velmi úspěšná a existuje několik modifikací Anti-Trust ranku – v českém prostředí je velmi populární tzv. Bad Rank [24]. Během experimentů, při kterých byla srovnávána účinnost metody Anti-Trust rank a Trust-Rank byly naměřeny následující hodnoty [23]:



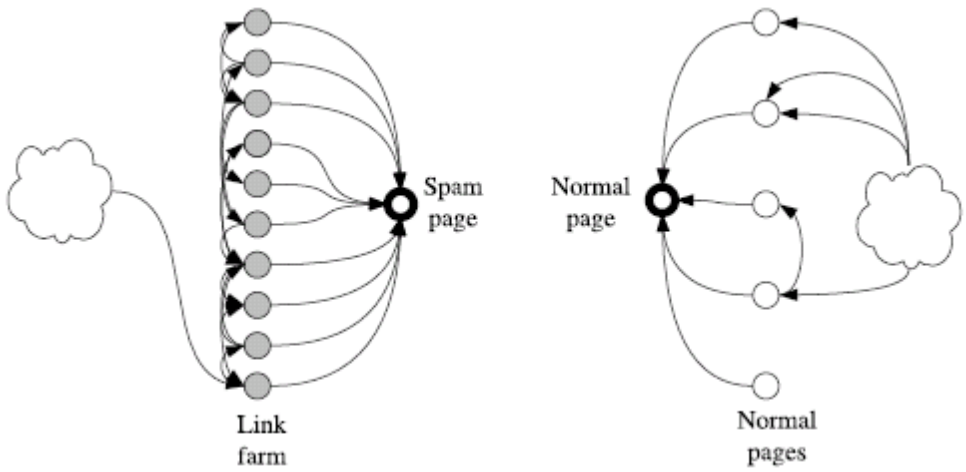
Obrázek 3-5 počet spamových stránek podle použitých metod s různými parametry

Tento graf uvádí počet spamových stránek nalezených pomocí Trust Ranku s množinou seed 40 stránek proti Anti-Trust ranku se stejně množinou 40, 80 a 120 stránek. Osa x představuje počet stránek, na které byla metoda aplikována, osa y znázorňuje množství spamových stránek objevených danou metodou. Všeobecně lze říci, že metoda Anti-Trust rank je úspěšnější a to především díky přesnějšímu určení pravděpodobnosti, že daná stránka je spam.

Výhodou nejen metody Anti-Trust rank ale Trust-Rank je fakt, že se dají použít nejen k analýze SEO spamu ale i k analýze obsahové podobnosti stránek. Lze například určit seed stránek s vysokým Page Rankem pro slovo „erotika“ a pomocí těchto metod určit další stránky, které rovněž spadají do této tematické kategorie. Další výhodou této metody je fakt, že je schopna odhalit spamové stránky s velkými PR, tedy většinou velmi komerční stránky, které jsou spravovány „seo firmami“. V současné době je v popředí studií možnost kombinace obou výše uvedených metod [23].

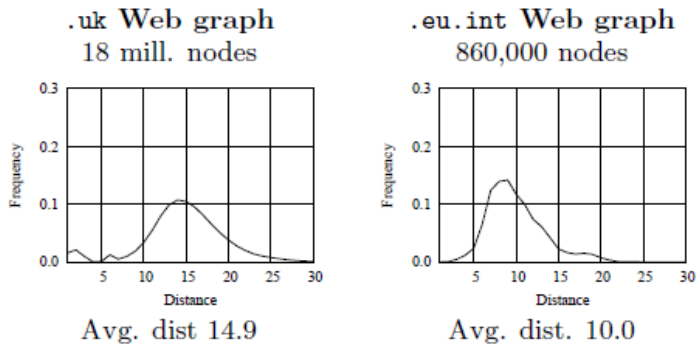
3.2.1.4 Truncated PageRank

Tato metoda je určitým vylepšením hodnocení Page Rank algoritmu ve snaze omezit působnost spamu a linkových farem. Vychází se zde z předpokladu, že spameři často využívají ranku velkého množství přímých odkazů k tomu, aby dosáhli vysokého hodnocení svých stránek. Toto se týká především linkových farem, které většinou vykazují následující linkovou strukturu [25]:



Obrázek 3-6 ukázka linkové farmy

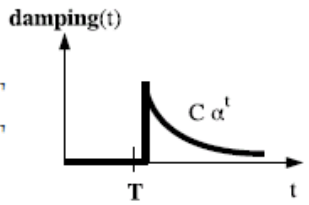
Budeme-li uvažovat, že stránka b odkazuje na stránku a, pak všechny stránky c odkazující na b svým způsobem rovněž podporují stránku a. Tyto stránky podporují stránku a s určitou vzdáleností t (což je nejmenší možný počet kliků, který je třeba provést, abychom se z podporující stránky dostali na stránku a). Všechny podporující stránky tvoří určitou strukturu, kterou lze znázornit grafem podle vzdálenosti, se kterou stránku podporují. Zde je ukázkový graf distribuce odkazů pro 2 sety náhodných stránek domén .uk a .eu.int [25]:



Obrázek 3-7 graf distribuce odkazů

Jak lze vidět, stránky mohou vykazovat určité anomálie v rozložení svých podporujících stránek. Z obrázku ukazujícího linkovou strukturu spamových stránek je patrné, že tyto stránky mají velký podíl své podpory získaný ze stránek s malou vzdáleností t. Tento fakt je pochopitelný, jelikož spameři, kteří shání zpětné odkazy na své stránky, se většinou starají pouze o kvalitu stránek, ze kterých odkazy nakupují. V praxi se měří většinou několik charakteristik typu PR, počet odchozích odkazů, množství stránek zasahující odkaz (pokud je kupovaný odkaz site-wide, tedy odkaz vyskytující se na každé stránce daného webu), ale již není prostor ani možnost hodnotit, jaké stránky na daný web odkazují, či jaké stránky odkazují na tyto stránky. Pokud se tedy utlumí přímý dopad nejbližších sousedů na hodnocení stránky, nejen že se sníží vliv nakupovaných odkazů, ale zároveň se eliminuje spousta jednoduchých linkových farem. Právě tuto myšlenku se snaží zavést Truncated PageRank pomocí modifikace damping vectoru (vektor útlumu vlivu stránky) ve vzorci výpočtu Page Ranku, dle následující funkce [25]:

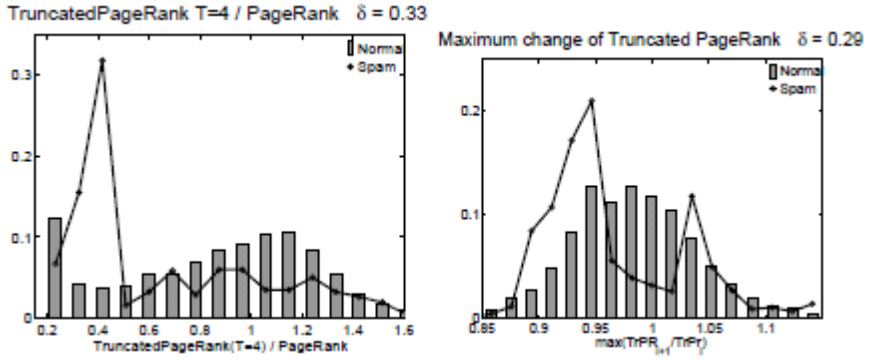
$$\text{damping}(t) = \begin{cases} 0 & t \leq T \\ C\alpha^t & t > T \end{cases}$$



Obrázek 3-8 funkce damping faktoru

Tato funkce omezuje vliv topologicky blízkých webů v grafu odkazujících stránek. V praxi to znamená, že podpora se vypočítává pouze od webů ve větší vzdálenosti než je T (konstanta, která se volí nejčastěji v rozmezí 1-4) [25].

Tato metoda je poměrně úspěšná. V praxi se spamová stránka pozná tak, že její Truncated PageRank je podstatně menší než standardní PageRank. Tato situace je zřetelná na následujícím obrázku vlevo [26]:



Obrázek 3-9 použití truncated PageRanku s různou konstantou T

Na tomto obrázku je patrný ostrý vrchol v malých vzdálenostech, který získají spamové stránky díky použití Truncated PageRanku. Zajímavé je rovněž sledovat poměr mezi Truncated PageRankem s distanční vzdáleností $T = i$ a distanční vzdáleností $T = i - 1$ (graf vpravo) – z těchto údajů lze rovněž určit s jakou pravděpodobností je stránka spamová. Při použití této metody lze odhalit 76,9% až 78% spamových stránek s pravděpodobnostní chybného označení 1,6% - 2,5% [26].

3.2.1.5 Spam rank

Spam rank je metoda, která vychází z poznatku, že existuje spoustu stránek, jejíchž PR (Page Rank) byl získán pomocí velkého množství neobjektivních odkazů směřujících na danou stránku. Tato metoda nerozlišuje, jestli byl daný odkaz získán přirozeně nebo uměle, ale zaměřuje se na jeho kvalitu. Tato metoda vychází z použití Personalized Page Rank algoritmu [17], který modifikuje následujícím způsobem:

$\boldsymbol{for\ all}\ Web\ pages\ i\ \boldsymbol{do}$

$vytvoř\ prázdné\ vektory\ \ Support_i$

$\boldsymbol{Fáze\ 1}$: $\boldsymbol{Generování\ podpůrců}$ ($\boldsymbol{supports}$)

$Do\ vektoru\ Support_i\ zařaď\ všechny\ stránky, které\ mají\ vysoký\ podíl\ na\ Page\ Rank\ i$

$\boldsymbol{Fáze\ 2}$: $\boldsymbol{Penalizace}$

$\boldsymbol{for\ all}\ Web\ pages\ i\ \boldsymbol{do}$

$přiděl\ Penalty_{\ i}\ v\ případě\ nestandardního\ chování\ PageRanku\ v\ Support_i$

$\boldsymbol{Fáze\ 3}$: $\boldsymbol{výpočet\ SpamRanku\ podle\ míry\ Penalties}$

Klíčová fáze v tomto algoritmu je ono určení nestandardního chování Page Ranku u stránek, které podporují uvažovanou stránku (supportem). Tato metoda vychází ze zkoumání sousedních podpůrných stránek za použití následujících metrik:

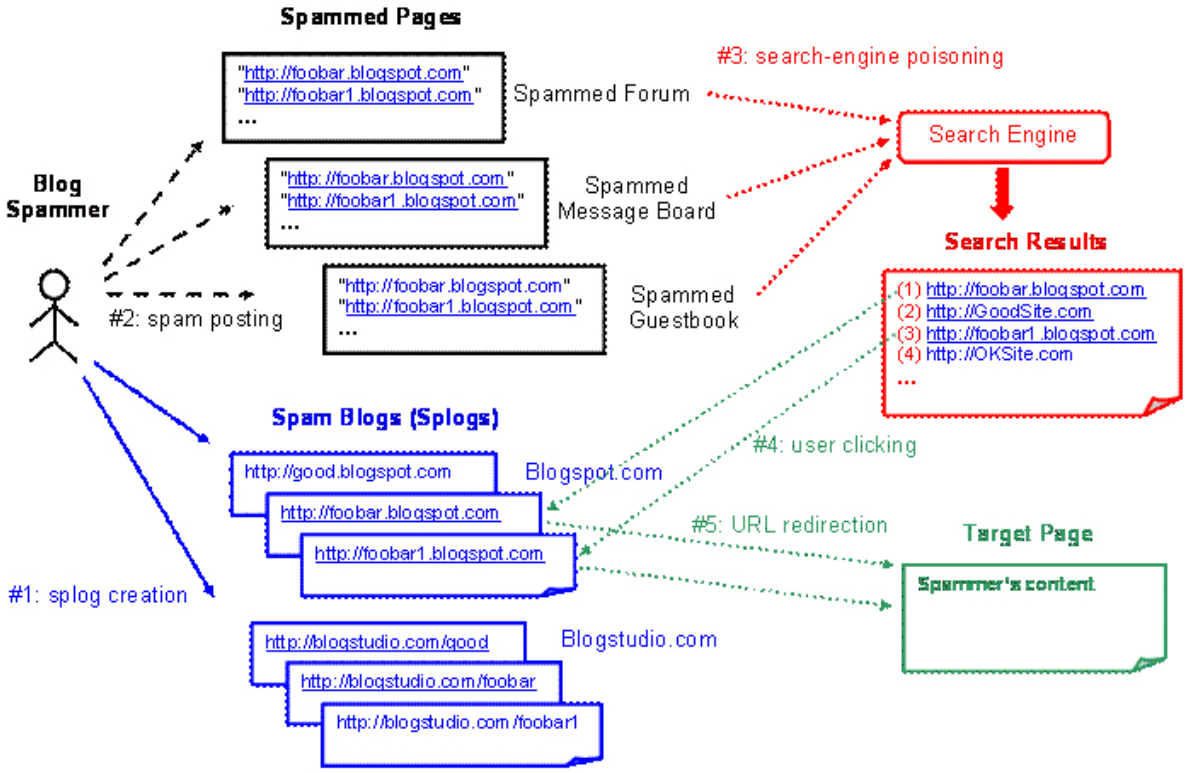
- Odkazující webové stránky vykazují podobný (většinou velmi nízký) Page Rank. V tomto případě se předpokládá, že rozličnost přirozených odkazů na internetu je vysoká – tedy na stránku budou odkazovat stránky s různými PR.

- Odkazující webové stránky vykazují podobnosti ve struktuře. Zde se předpokládá, že spameři mají příliš malý rozpočet na to, aby mohli tvořit opravdu unikátní stránky se zpětnými odkazy.

Síla této metody spočívá především v analýze důvěryhodnosti PR stránek, z nichž daná stránka čerpá. Vylepšená verze této metody počítá s histogramy PR – pokud jsou nápadně podobné, lze vypočítat kolik PR daná stránka dostala neoprávněně (Spam Rank). Tato metoda byla otestována na vzorku 31 miliónu stránek z německých domén a v daném vzorku odhalila spamovost 16,5% [18].

3.2.1.6 Strider Search Defender

Search Defender je projektem výzkumného týmu Strider, který je financován společností Microsoft. Tento tým vyvíjí sadu nástrojů pro automatickou detekci seo-spamu. Tyto nástroje se zaměřují především na odhalování doorway stránek, spamování ve veřejných fórech a vytváření spamových blogů – tzv. Splogs. Vychází se zde z velmi častého chování spamerů, které demonstruje následující obrázek [27]:



Obrázek 3-10 výchozí chování spamerů podle výzkumného týmu Strider

Předpokládá se, že spameři získávají zpětné odkazy tak, že používají automatizované programy pro vkládání náhodných příspěvků do otevřených fór s odkazem na cílenou stránku. Těchto programů se v současné době používá hned několik a jedná se o poměrně snadno dostupné komerční aplikace. Tomuto spamování se říká comment spamming. Spameři rovněž velmi často používají doorway stránky jako cíle comment spammingu. Tímto lze zobrazit url, které vypadá důvěryhodněji a lze

získat více návštěvnosti pomocí přesměrování. Dalším způsobem spamu jsou seo-účelové blogy (Splogs) podporující cílovou stránku (ať už formou zpětných odkazů nebo pomocí přesměrování).

Search Defender používá pro odhalení těchto podvodných praktik 2 hlavní nástroje. V první fázi je použit tzv. Spam Hunter. Tento program vychází z ručně naplněného listu potvrzených spamových stránek. Tyto stránky postupně vkládá jako argument dotazu link: čímž zjistí stránky, které odkazují na danou spamovou stránku. Tento proces opakuje pro všechny stránky v listu tak dlouho, dokud nárůst nových odkazujících stránek je větší než zadaná konstanta. Takto lze získat velké množství potenciálních spamových stránek, u kterých však nemůžeme s jistotou říci, jestli se opravdu jedná o spam či nikoliv.

Pro toto rozlišení používá Search Defender další nástroj, který se jmenuje Strider URL Tracer. Vstupem tohoto nástroje je množina potenciálních spamových stránek naplněná Spam Hunterem. Klíčovou funkcí Strider URL Trackeru je tzv. Top Domain view. Tato funkce simuluje reálného návštěvníka na stránkách (tedy používá některý z aktuálních webových prohlížečů), aby mohla detekovat případná přesměrování. Top Domain view navštíví každou stránku z URL listu a zaznamenává, na které další stránky se lze z navštívené stránky dostat či na které byl přesměrován. Vzhledem k faktu, že lze předpokládat, že URL list obsahuje potenciálně spamové stránky, můžeme seřadit URL list podle počtu výše zaznamenaných stránek generujících odkaz či přesměrování. Takto lze získat seznam domén, které jsou seřazeny podle pravděpodobnosti, že se jedná o spam. Tento nástroj je volně ke stažení na stránkách výzkumného týmu Strider [28].

Search Defender má dvě vlastnosti, které aktivity spamerů obrací proti nim samotným:

1) Čím více spamu směřuje na některou stránku, tím snazší je pro Spam Henter tuto stránku najít. Jakmile je spamové fórum ověřeno, stane se z něj „HoneyForum“, které může být použito k zachycení nových spam adres.
2) Čím více doorway stránek či splogů spamer vytvoří, tím vyšší pozici odkazovaná doména získá v Top Domain view a je tedy předurčena k podrobnějšímu vyšetření.

V ideálním případě je Spam Hunter schopen identifikovat nové spamové URL a upozornit na ně vyhledávače (které tyto URL budou ignorovat) ještě před tím, než daná spamerská aktivita negativně ovlivní kvalitu výsledků vyhledávání. Toto je možné vzhledem k faktu, že existuje prodleva mezi zavedením spamu a ovlivněním SERPu.

3.2.2 Metody využívající analýzu on-page faktorů

Jelikož on-page faktory jsou nejjednodušší věcí, kterou spamer může ovlivňovat, všeobecnou tendencí moderních vyhledávačů je snižovat význam on-page faktorů při hodnocení pořadí ve výsledcích vyhledávání. Na druhou stranu jsou to však právě on-page faktory, které významně

napovídají, jestli je stránka relevantní vzhledem k zadanému dotazu. Proto existuje mnoho způsobů, jak vyhledávače poznají, jestli je stránka „přeoptimalizovaná"

Nejjednodušším způsobem je měření density (tedy procentuálního zastoupení) slov. Jestliže je některé slovo zastoupeno na stránce ve zvýšeném množství, je pravděpodobné, že tato stránka je souvztažná k danému slovu. Pokud však tato densita přesáhne určitý práh, pak se pravděpodobně jedná o tzv. „seo text“, tedy text s nepřirozeným a často velmi umělým výskytem tohoto slova za účelem získání vyšší relevance stránky. Vyhledávače tedy mohou při překročení prahu density naopak stránky penalizovat a tímto snižovat účinnost spamerské metody vkládání klíčových slov – tzv. „keywords stuffing“.

Vzhledem k faktu, že většina seo optimalizátorů počítají s výše uvedenou metodou, spamerské stránky často respektují onu „rozumnou hranici“ density slov a předchozí metoda je tedy neúčinná. Proto se vyhledávače snaží detekovat jakékoliv anomálie v on-page faktorech stránky oproti přirozenému obsahově hodnotnému textu. Jednou z možných metod sledování obsahové hodnoty a věrohodnosti textu patří získávání informací o stránce pomocí detekce frází [29]. Tato metoda je založena na poznání, že výskyt některých frází v dokumentu predikuje výskyt dalších frází. Například fráze „president spojených států“ predikuje výskyt fráze „bílí dům“. Metoda pracuje pomocí tohoto principu (ve zjednodušené podobě):

- Vyhledávač si během indexování vytvoří tabulku frází, kde pomyslná osa x a y obsahuje všechny časté fráze (tedy ty, které se vyhledávají) vyskytující se na stránce
- Porovnává se počet výskytů 2 nezávislých frází v dalších tabulkách jiných stránek
- Jestliže je nalezeno dostatečné množství spoluvýskytů, je pravděpodobné, že fráze na ose x predikuje frázi na ose y a je tedy v průniku těchto dvou frází v tabulce zapsána příslušná hodnota
- V rámci porovnávání relevance k vyhledávané frázi se váží celý řádek tabulky souvztažný k danému slovu

Další metody detekce SEO spamu pomocí analýzy on-page faktorů vychází z poznatku, že mnoho linkových farem používá tzv. patičky – tedy místa umístěná většinou na spodku stránky, která obsahují zpětné odkazy na další weby. Smysl těchto patiček je pouze v tom, že spameři takto získávají zpětné odkazy na další webové stránky. Metody detekující tyto patičky vychází z několika principů:

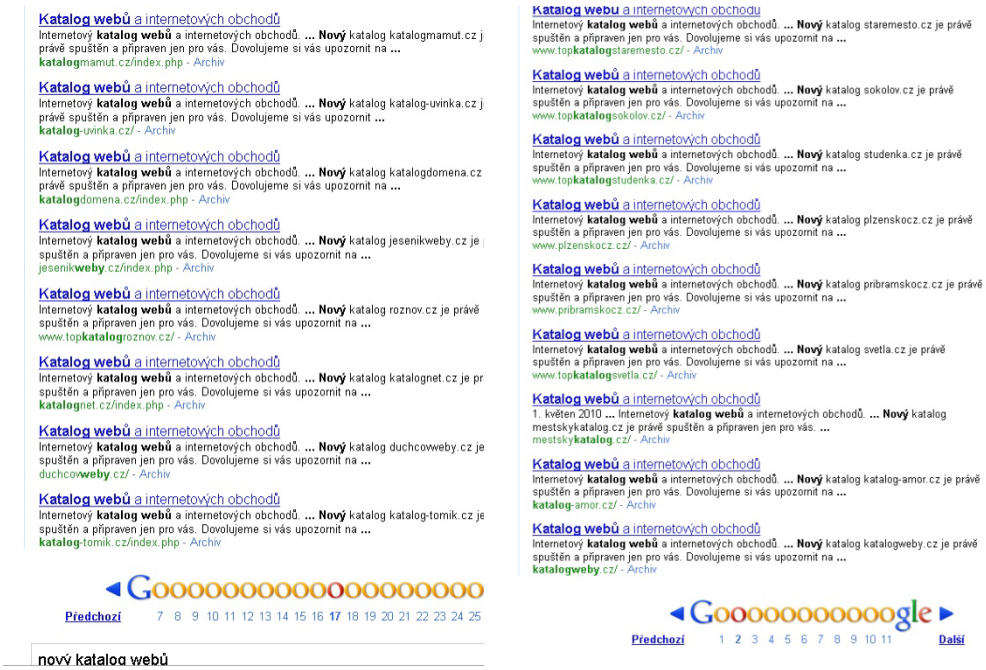
- Analýza okolí zpětných odkazů – jestliže je mnoho odkazů odkazující na externí domény v těsné blízkosti u sebe, pak se s vysokou pravděpodobností jedná o patičku.
- Analýza tematické příbuznosti odkazů vzhledem k stránce – metoda vychází z poznatku, že spameři často odkazují na jakékoliv weby, bez ohledu, jestli jsou nebo nejsou hodnotné pro návštěvníka daných stránek. Pokud tedy návštěvník prohlíží stránku, která se zabývá pravidly českého pravopisu, pak jsou pro něj bezcenné odkazy typu „levné bazény“ bezcenné a na této stránce nemají co dělat.

- Analýza výskytu tzv. jedovatých slov. Tato metoda vychází z množiny slov, které se často pojí s výskytem patiček a dalších spamerských metod. Jedná se například o slova odkazy, další odkazy, linky, m-link, atd. Pokud se v blízkosti těchto slov nachází zvýšené množství odkazů na externí stránky, pak se s vysokou pravděpodobností jedná o spamerský počin.
- Detekce automatických odkazových systémů. Odkazové systémy fungují na principu, že se do nich administrátor zaregistruje (občas za to i zaplatí) a dostane kus php kódu, který vloží na jeho stránky. Tento kód pak již komunikuje s provozovatelem systému a ten porovnává vhodné kandidáty na výměnu zpětných odkazů. Automaticky se pak na stránkách obsažených v systému generují zpětné odkazy s cílem zvýšit hodnocení členů v tomto systému. Tento systém je účinný především díky tomu, že weby často obsahují různé odkazy na každé své stránce – takto lze získat odkazový profil blízký ideálnímu přirozenému rozdělení, tedy že množství webů, které na danou stránku odkazuje je blízké počtu odkazujících stránek (tento systém potírá tzv. site-wide odkazy). Nejznámějším příkladem takovéhoto výměnného systému v českém internetovém prostředí je tzv. Expert lišta. Jedná se o placený nástroj, který se snaží vyměňovat tematické odkazy. Tyto nástroje jsou poměrně účinné a většina vyhledávačů je zakazuje [30].
- Výskyt podpisů „SEO firem“. Jedná se o doplňkovou metodu, která analyzuje, jestli se na webové stránce nevyskytuje popis typu „SEO made by profiSEO“ nejčastěji se zpětným odkazem na danou společnost zabývající se SEO optimalizací. Odhalení, že stránku optimalizovala nějaká společnost, samo o sobě není škodlivým faktem (může se jednat o čestnou optimalizaci, která naopak vyhledávačům pomáhá správně rozlišit, čím se stránka zabývá). Na druhou stranu při zjištění, že se na stránce objevil podpis „SEO firmy“, může napovědět vyhledávači, že by měl stránku podrobněji sledovat. Příkladem takovéhoto sledování může být porovnávání histogramů linkového portfolia (tedy kolik a jakých stránek na danou stránku odkazuje). Jestliže například na stránku odkazovalo dlouhodobě 500 stránek, z nichž většina byla z katalogů, a po vložení podpisu SEO společnosti začne množství odkazujících stránek prudce stoupat, pak je zjevné, že se jedná o umělé odkazy, které stránce zařídila ona „SEO firma“.

Jak lze vidět, on-page faktory zle sledovat z mnoha úhlů. Všeobecně platí, že jakákoliv anomálie, která z pohledu samotného uživatele nepřináší žádný užitek, je nežádoucí a měla by tedy být penalizována. V současné době lze totiž pozorovat, že spousta i kvalitních webů je „zohyzděna“ seo-lištami, účelově napsanými seo-nadpisy a dalšími pro uživatele otravnými prvky.

3.3 Budoucnost detekce SEO spamu

Spousta existujících metod detekce SEO spamu trpí tím, že seo-spameři také nespí a neustále zdokonalují své spamové techniky (například systémy pro automatickou výměnu odkazů přibývají velkým tempem a jsou stále rozmanitější a tedy obtížněji detekovatelné). Dalším problémem je fakt, že některé spamerské techniky je občas běžnými metodami i nemožné účinně detekovat (například nákupy zpětných odkazů). Rovněž lze říci, že spameři mohou být v některých případech o krok dopředu, jestliže vyvinou novou metodu, proti které se vyhledávač neumí bránit – příkladem zmanipulování googlu je například zadání dotazu „nový katalog webů“ který od 2. do 17. stránky generuje nápadně podobné výsledky vyhledávání – viz následující obrázky:



Obrázek 3-11 ***Demonstrace „ovládnutí“ googlu. Pořízeno 9.5.2010***

Všeobecně řečeno spameři se snaží vyzrát nad vyhledávači a vývojáři vyhledávačů nad spamery. Jelikož je však spamerů podstatně více, jedná se o poměrně nerovný souboj. Proto za budoucnost rozpoznávání seo-spamu je považováno strojové učení. Jedná se o navržení klasifikátoru, který se bude schopen učit od samotných spamerů a efektivně rozpoznávat seo-spam. Pro toto použití se rovněž uvažuje i nasazení inteligentních agentů [31] [32] [33], kteří budou použity místo jednoúčelových webcrawlerů. Každý z těchto agentů bude reflektovat určitou roli uživatele na internetu. Bude tedy navštěvovat množinu svých oblíbených či doporučených stránek a sledovat jejich změny, zároveň však bude komunikovat s dalšími agenty a vyměňovat si získané zkušenosti a poznatky. Takto se agenti budou učit, které stránky jsou kvalitní a čemu se naopak vyvarovat. Celkový provoz agentů budou řídit další agenti a všeobecně mezi agenty bude určitá funkční hierarchie, kterou demonstruje následující obrázek [31]:

Obrázek 3-12 nasazení agentů ve vyhledávačích nové generace

Strojové učení a umělá inteligence je bezesporu perspektivním směrem nejen boje proti seo-spamu, ale všeobecně pro zlepšení kvality vyhledávání. V současné době již mnohé vyhledávače používají fuzzy logiku k automatické kalibraci konstant ovlivňujících různé faktory hodnocení stránek – viz další kapitola.

4 Měření kvality výsledků vyhledávání

Výsledky vyhledávání (SERP) poskytuje každý vyhledávač jiné, abychom byli schopni rozlišit, který vyhledávač je lepší, můžeme se na tuto problematiku dívat z 2 hledisek:

1. Komerční úspěch vyhledávače – ten je však ovlivněn mnoha faktory, které nesouvisí s opravdovou kvalitou jeho výsledků (například investice do reklamy, historický vývoj trhu, uživatelská přívětivost, množství reklamy ve vyhledávání, atd.)
2. Kvantitativně porovnávat výsledky vyhledávání s tzv. kalibračními daty

První bod lze měřit poměrně přesně (například z globálních statistik některých monitorovacích služeb měřících referer). Zde jsou statistiky podílů vyhledávačů v českém internetovém prostředí [34]:

Podíl vyhledávačů v refereru [historie]					
Kategorie	Vyhledávače				
1. Auto, moto	Seznam (46.90 %)	Seznam (Fulltext) (28.48 %)	Google.CZ (10.22 %)	Google.SK (3.77 %)	Firmy.cz (Seznam) (2.45 %)
2. Cestování	Seznam (Fulltext) (45.95 %)	Google.CZ (21.57 %)	Google.SK (8.55 %)	Google (6.89 %)	Google.DE (3.06 %)
3. Časopisy	Google.CZ (39.54 %)	Seznam (Fulltext) (38.10 %)	Google.SK (10.81 %)	Google (4.67 %)	Centrum (Google) (2.05 %)
4. Divadlo	Seznam (Fulltext) (36.40 %)	Google.CZ (32.17 %)	Google.SK (18.60 %)	Google (7.39 %)	Centrum (Google) (2.05 %)
5. Ekonomika, finance	Seznam (Fulltext) (49.88 %)	Google.CZ (34.62 %)	Google.SK (5.04 %)	Google (4.55 %)	Centrum (Google) (2.08 %)
6. Film	Seznam (Fulltext) (46.48 %)	Google.CZ (32.79 %)	Google.SK (11.91 %)	Google (4.56 %)	Seznam katalog (1.61 %)
7. Fotografie	Seznam (Fulltext) (38.95 %)	Google.CZ (26.80 %)	Google.SK (10.21 %)	Google (8.87 %)	Google.PL (2.49 %)
8. GSM	Seznam (Fulltext) (55.48 %)	Google.CZ (19.78 %)	Google.SK (8.55 %)	Zbozi.cz (Seznam) (8.34 %)	Google (3.41 %)
9. Hardware	Google.CZ (36.81 %)	Seznam (Fulltext) (30.37 %)	Zbozi.cz (Seznam) (17.70 %)	Google (4.32 %)	Google.SK (4.09 %)
10. Historie	Seznam (Fulltext) (39.61 %)	Google.CZ (28.50 %)	Google.SK (21.42 %)	Google (4.95 %)	Zbozi.cz (Seznam) (0.94 %)
11. Hudba	Seznam (Fulltext) (44.46 %)	Google.CZ (26.29 %)	Google.SK (18.74 %)	Google (5.78 %)	Centrum (Google) (0.90 %)
12. Chovatelství	Seznam (Fulltext) (56.34 %)	Google.CZ (24.62 %)	Google.SK (8.43 %)	Google (3.06 %)	Zbozi.cz (Seznam) (2.88 %)
13. Instituce	Seznam (Fulltext) (59.83 %)	Google.CZ (28.88 %)	Google.SK (3.93 %)	Google (3.09 %)	Centrum (Google) (1.82 %)
14. Internet	Seznam (Fulltext) (40.06 %)	Google.CZ (18.61 %)	Google (13.31 %)	Google.SK (9.75 %)	Google.CO.UK (3.26 %)
15. Inzerce, reklama	Seznam (Fulltext) (55.40 %)	Google.CZ (22.31 %)	Google.SK (7.62 %)	Google.PL (5.33 %)	Google (3.33 %)
16. Komerční stránky	Seznam (Fulltext) (48.01 %)	Google.CZ (21.94 %)	Zbozi.cz (Seznam) (13.99 %)	Google.SK (5.51 %)	Firmy.cz (Seznam) (3.65 %)
17. Koníčky, hobby	Seznam (Fulltext) (60.04 %)	Google.CZ (23.74 %)	Google.SK (7.14 %)	Google (3.14 %)	Centrum (Google) (1.64 %)
18. Neuvedeno	Seznam (Fulltext) (51.46 %)	Google.CZ (24.84 %)	Google.SK (10.69 %)	Google (3.60 %)	Zbozi.cz (Seznam) (3.24 %)
19. Obce, města	Seznam (Fulltext) (52.70 %)	Google.CZ (32.85 %)	Google.SK (5.46 %)	Google (4.08 %)	Centrum (Google) (2.02 %)
20. Osobní stránky	Seznam (Fulltext) (50.91 %)	Google.CZ (26.31 %)	Google.SK (11.04 %)	Google (5.26 %)	Centrum (Google) (1.10 %)
21. Počítačové hry	Seznam (Fulltext) (55.12 %)	Google.CZ (24.62 %)	Google.SK (8.73 %)	Google (5.16 %)	Seznam (0.95 %)
22. Průmysl	Seznam (Fulltext) (55.92 %)	Google.CZ (24.06 %)	Google.SK (5.72 %)	Firmy.cz (Seznam) (3.78 %)	Google (3.27 %)
23. Rádia	Seznam (Fulltext) (62.17 %)	Google.CZ (22.95 %)	Google.SK (5.83 %)	Google (3.68 %)	Centrum (Google) (1.85 %)
24. Reality	Seznam (Fulltext) (57.34 %)	Google.CZ (28.72 %)	Firmy.cz (Seznam) (5.03 %)	Google (3.12 %)	Google.SK (2.20 %)
25. Software	Seznam (Fulltext) (49.31 %)	Google.CZ (29.97 %)	Google.SK (8.63 %)	Google (5.50 %)	Centrum (1.29 %)
26. Sport	Seznam (Fulltext) (53.91 %)	Google.CZ (21.92 %)	Google.SK (9.44 %)	Google (4.61 %)	Zbozi.cz (Seznam) (3.09 %)
27. Stavebnictví	Seznam (Fulltext) (49.82 %)	Google.CZ (29.65 %)	Google.SK (5.47 %)	Zbozi.cz (Seznam) (5.38 %)	Google (2.85 %)
28. Státní správa	Seznam (Fulltext) (45.88 %)	Google.CZ (41.66 %)	Google (4.54 %)	Google.SK (2.55 %)	Firmy.cz (Seznam) (1.96 %)
29. Školství	Seznam (Fulltext) (41.21 %)	Google.CZ (33.58 %)	Google.SK (12.79 %)	Google (5.46 %)	Google.PL (2.20 %)
30. Umění	Seznam (Fulltext) (48.92 %)	Google.CZ (28.98 %)	Google (6.94 %)	Google.SK (4.83 %)	Google.DE (3.84 %)
31. Věda, technika	Seznam (Fulltext) (39.40 %)	Google.CZ (30.25 %)	Google.SK (19.75 %)	Google (6.30 %)	Centrum (Google) (1.01 %)
32. Virtuální obchody	Seznam (Fulltext) (41.37 %)	Google.CZ (27.12 %)	Zbozi.cz (Seznam) (16.03 %)	Google.SK (8.73 %)	Google (2.91 %)
33. Weblogy	Google.SK (33.99 %)	Seznam (Fulltext) (30.65 %)	Google.CZ (25.45 %)	Google (6.42 %)	Centrum (Google) (0.51 %)
34. Webziny	Seznam (Fulltext) (51.69 %)	Google.CZ (33.24 %)	Google.SK (6.50 %)	Google (4.88 %)	Centrum (Google) (1.55 %)
35. Zábava	Seznam (Fulltext) (55.62 %)	Google.CZ (19.86 %)	Google.SK (14.43 %)	Google (3.73 %)	Seznam katalog (1.81 %)
36. Zdraví	Seznam (Fulltext) (56.73 %)	Google.CZ (24.88 %)	Google.SK (8.26 %)	Google (3.63 %)	Zbozi.cz (Seznam) (1.76 %)
37. Zemědělství	Seznam (Fulltext) (46.65 %)	Google.CZ (19.43 %)	Google (8.42 %)	Google.SK (8.09 %)	Google.PL (6.42 %)
38. Zpravodajství	Seznam (Fulltext) (38.14 %)	Google.CZ (26.88 %)	Seznam katalog (11.50 %)	Google.SK (9.12 %)	Google (4.46 %)
39. Životní styl	Seznam (Fulltext) (51.10 %)	Google.CZ (26.80 %)	Google.SK (11.11 %)	Google (3.80 %)	Zbozi.cz (Seznam) (1.99 %)

Popis: výsledek je brán z dat za včerejšek u všech stránek, které měří referer, kromě kategorie Erotika, Anglicky a statistiky samotných serverů TOPlist a Caramba!.
Započítány jsou servery, které přinesly alespoň 10 prokliků. Vypsáno je prvních 5. Zbytek do 100% jsou ostatní servery, které splnily podmínku.

Obrázek 4-1 podíl vyhledávačů na českém trhu

Z těchto statistik je evidentní, že trh s vyhledáváním v českém internetu ovládá vyhledávač seznam.cz a google, nicméně nelze jednoznačně říci, který vyhledávač poskytuje relevantnější výsledky vyhledávání a je tedy „lepší“. Tato otázka je svým způsobem i velmi těžko definovatelná, jelikož nelze říci, které dotazy budou pro srovnání použity a kdo bude určovat onu relevanci. Problém definice porovnávání dotazů vyplívá z faktu, že množina možných dotazů je v podstatě nekonečná. Co se týče samotné relevance stránek obsažených v SERPu, narážíme zde na fakt, že ohodnocení této relevance je velmi subjektivní a záleží na sociálních faktorech uživatele, předešlých zkušenostech, IT gramotnosti, momentální náladě, časové tísni, atd. Dále zde narážíme na problém, co sledovat? Jestli jen kvalitu samotných stránek v SERPu, rychlost hledání, kvalitu snippetů, nebo jiné parametry.

4.1 Tvorba kalibračních dat

Výše uvedené problémy lze řešit pomocí dostatečně statisticky přesných vzorků dotazů ohodnocených reálnými uživateli z různých sociodemografických vrstev. V podstatě se vezme určité množství náhodně vybraných dotazů (zde se může přihlížet i k jejich statistickému objemu vyhledávání) a tyto dotazy jsou předkládány před dostatečné množství uživatelů různého pohlaví, z různých věkových skupin, zájmů, atd., kteří procházejí SERP k danému dotazu a ohodnocují relevanci nalezených stránek. Pro toto ohodnocení se používá většinou zařazení stránky do určité kategorie podle subjektivní relevance vzhledem k danému dotazu. Zde je možná ukázka těchto kategorií:

- **Vitální** – stránka je jedinečná (často oficiální), dotaz má jasnou interpretaci
  Příklad: dotaz „facebook“ | stránka facebook.com
- **Užitečná** – stránka je velmi uspokojující, poskytuje vyčerpávající informace vzhledem k dotazu, lze ji považovat za důvěryhodný zdroj
  Příklad: dotaz „angličtina“ | stránka helpforenglish.cz
- **Dobrý výsledek** – průměrná stránka, relevantní vzhledem k výsledku
  Příklad: dotaz „harry potter“ | stránka knihy.cz/prodej/harry-potter
- **Špatný výsledek** – stránka je sice relevantní k tématu, nicméně obsahuje příliš obecné nebo staré informace, může být pro uživatele zavádějící
  Příklad: dotaz „brno“ | stránka www.classicshow.cz (Classic Show Brno 2010)
- **Off-topic** – stránka sice obsahuje hledaný výraz, nicméně buď ve špatném kontextu nebo účelově
  Příklad: dotaz „přepočet kurzů“ | stránka http://www.boraborahotel.com/Eden-Beach-hotel-Bora-Bora-cz/Jitka_a_zraloci.html (Jitka a žraloci)

Takto se kalibruje vzorek dotazů (řádově tisíce). Přičemž se uvažují i další aspekty jako porozumění dotazu (například při odborném dotazu), vztah k dotazu (muž x žena), atd.

4.2 Použití kalibračních dat

Výše zmíněné kalibrační data jsou pro vyhledávače velmi důležité. Nejen že takto jsou schopni srovnat kvalitu svého vyhledávání s konkurencí, ale mohou odhadovat úspěšnost nové metody použité při hodnocení stránek ještě dříve, než tuto metodu pustí do ostrého provozu. Toto řídícím pracovníkům vývojových týmů vyhledávače umožňuje relevantně se rozhodovat na základě reálných dat. Dalším velkou výhodou použití kalibračních dat je fakt, že manageři vyhledávače včasně zjistí, jak se zlepšila konkurence a mohou na to pružně reagovat. Většinou se kalibrací dat zabývají externí

kalibrátoři, tudíž se přenáší značná část práce a vývojový tým vyhledávače se tedy může plně soustředit na svou práci. Navíc lze v rámci kalibrace rozpoznat spoustu spamových stránek a tyto označené stránky použít jako seed pro metody typu AntiTrust Rank (v českém prostředí známá modifikace též modifikace Bad Rank).

4.2.1 Automatické nastavování parametrů vyhledávače pomocí kalibračních dat

Jedním z dalších významů kalibrace je možnost objektivního nastavování parametrů používaných při hodnocení stránek v metodách vyhledávače. Jedná se například o konstanty typu „zdravá densita slov“, kdy je velmi obtížné (ne-li nemožné) určit přesnou hodnotu této konstanty. Automatická kalibrace umožňuje sledovat vývoj kvality v závislosti na nastavení určitých vah (což můžou být několika rozměrné grafy funkcí) a hledat vrcholy těchto „grafů“ (nejlepší kvalitu výsledků). Tímto lze najít hodnoty parametrů takové, že výsledky vyhledávání jsou v rámci použitých metod nejkvalitnější.

5 Návrh vlastní metody proti SEO spamu

Při návrhu metody, která by byla využitelná vyhledávači při boji proti SEO spamu jsem vycházel z několika postulátů:

- Kvalita vyhledávání závisí na kvalitě uživatelova dotazu
- Uživatel při formulaci svého dotazu často využije dotaz z nabídky našeptávače
- Výsledek vyhledávání ze zadaného dotazu nemusí přesně odpovídat představě uživatele, a proto uživatel dotaz mění
- Lze předpokládat, že ve výsledcích vyhledávání se objevují nekvalitní stránky, které podporuje SEO spam
- Spameři používají SEO spam především na lukrativní klíčová slova, které nabízí našeptávač
- Oborově příbuzné stránky na sebe odkazují.
- Tyto odkazy mají větší vypovídající hodnotu, pokud odkazující stránka má přímou souvislost s dotazem, který uživatel zadal

- Výsledky vyhledávání obsahují z větší části stránky, které nepoužívají SEO spam

Snahou navrhované metody je odfiltrovat z výsledků vyhledávání stránky, které by se do SERPu dostali pomocí SEO spamu, a zároveň zvýšit kvalitu vyhledávání. Jelikož off-page faktory nemohou spameři snadno ovlivňovat, je tato metoda založena právě na analýze těchto faktorů.

5.1 Princip navrhované metody

Navrhovanou metodu jsem nazval Korelace výsledků vyhledávání. Tato metoda totiž vychází z funkčního vyhledávače, jehož SERP je následně modifikován tak, aby bylo dosaženo co nejrelevantnějších výsledků vyhledávání a aby byly sníženy pozice stránek vyhnaných SEO spamem. Tato metoda rovněž používá dalších výrazů z našeptávače pro získání dalších SERPů. Stránky z těchto výsledků vyhledávání jsou pak analyzovány pro zjištění závislosti mezi získanými stránkami. Z těchto závislostí je odvozen nový SERP, který odpovídá stanoveným postulátům.

Postup této metody je následující:

1) Uvažujeme množinu m_1 jejíž velikost $|m| = n$
2) Vezmi n stránek z výsledků vyhledávání pro daný uživatelův dotaz a tyto stránky ulož do množiny m_1
3) Uvažujme další množinu m_2, jejíž velikost $|m_2| = x$ (způsob výpočtu této velikosti bude popsán v další kapitole).
4) Vezmi x stránek z výsledků vyhledávání pro nejbližší další dotaz příbuzný dotazu, který uživatel zadal (vzato z našeptávače, respektive ze statistik hledanosti dotazů) a tyto stránky ulož do množiny m_2. Pokud se stránka vyskytuje v některé z předchozích množin m, je přeskočena.
5) opakuje bod 3 a 4 pro m_3, m_4, atd. dokud $x = |m_i| > 0$
6) každá stránka náležící množině $m_1..m_i$ dostane ohodnocení rp (relevancy points), které odpovídá velikosti množiny, do které stránka patří, mínus pozice této stránky v SERPu náležícím této množině.

$$rp = |m_x| - pos(s);\ \ rp \in m_x, pos - funkce\ pozice\ v\ SERP$$

7) každé stránce s je vypočítán trp (total relevancy points) – ten je vypočítán jako rp dané stránky . počáteční *boost* (kalibrační konstanta) + suma rp všech stránek odkazujících na tuto stránku. Vlastnost, kdy stránka $s_1$odkazuje na stránku s_2 znázorňujeme pomocí $\rightarrow$.

$$trp(s) = rp(s).boost + \sum rp(s_j);\ s_j \in m_1 \ldots m_i;\ s_j \rightarrow s$$

8) výsledný SERP je vytvořen seřazením všech stránek podle velikosti trp (vyšší trp = vyšší pozice v SERPu), stránky se stejným trp jsou seřazeny podle velikosti jejich rp/i
9) nový SERP je uložen do cashe, aby jej nebylo nutné opět počítat pro další instanci zadaného dotazu

Tato metoda je velmi variabilní a poskytuje mnoho prostoru pro modifikace. Největší výzvou je způsob výpočtu velikosti množin m, jelikož z těchto množin je přímo odvozen vliv dalších stránek, které se analyzují. Dalšími možnými modifikacemi jsou následující:

- eliminace/povolení opakování stránky v rámci více množin. V případě povolení opakování stránek je nutné po bodu 6 provést sumu trp dané stránky (z jednotlivých množin m)
- zohlednění množství podstránek, kterými daná stránka odkazuje. Tato modifikace by znamenala následující úpravu vzorce v bodu 6: kdy trp = suma rp*f(s),

$$trp(s) = rp(s)\,.boost + \sum rp(s_j).f(|s_j|);\ s_j \in m_1 \ldots m_i;\ s_j \rightarrow s$$

  kde f = korelační funkce, $|s_j|$ = počet stránek odkazujících na danou stránku
- zahrnutí tzv. hubů. Hub je stránka, na kterou často odkazují stránky z nalezených množin m. Lze předpokládat, že existují stránky, které jsou kvalitním zdrojem informací k dané tématice (proto je na tyto stránky často odkazováno), nicméně tyto stránky nejsou obsaženy v nalezených množinách m (například proto, že přímo neobsahují vyhledávaný dotaz). Tyto stránky jsou však kvalitním zdrojem rp, proto by se v této modifikaci při výpočtu trp zohledňovaly i zpětné odkazy z těchto hubů.
- zohlednění množství odkazujících podstránek z hubu
- vliv velikosti stránky pro stanovení počátečního rp. Tato modifikace je obzvláště vhodná, pokud zahrneme zohlednění množství odkazujících podstránek při výpočtu trp.

5.2 Návrh způsobu výpočtu velikosti množin m

Množiny $m_1 \ldots m_i$ zahrnují stránky, které se zobrazí v nemodifikovaném vyhledávači na daný výraz – nejprve původní, potom nejbližší další z našeptávače atd. Lze říci, že s rostoucím řádem množiny i klesá relevance výrazu, jelikož se již vzdaluje od dotazu, který původně uživatel zadal. Uvažujme referenční vyhledávač, ze kterého budeme skládat tyto množiny m, seznam.cz. Pokud budeme vycházet z volně dostupných informací o vyhledávanosti jednotlivých výrazů (z těch v podstatě vychází našeptávač tohoto vyhledávače), můžeme se setkat se 2mi případy:

- Hledaný řetězec je poměrně unikátní, a proto nemá zápis podobných výrazů ve statistikách hledanosti dotazu (např. http://search.seznam.cz/stats?collocation=kovoreal&submit=Vyhledat+Seznamem)
- Hledaný řetězec má zápis několika podobných výrazů a jejich hledanosti ve statistikách hledanosti dotazu (např. http://search.seznam.cz/stats?collocation=hry&submit=Vyhledat+Seznamem)

V prvním případě nemá smysl tvořit další řády množiny m. Metoda bude pracovat pouze s instancí prvního řádu této množiny. V druhém případě již můžeme odvodit další skupiny stránek, které jsou relevantní k danému tématu a jejichž analýza zpětných odkazů může zvýšit kvalitu vyhledávání.

Nyní musíme zohlednit fakt, že jednotlivé našeptávané výrazy (nejhledanější dotazy obsahující původní řetězec) jsou vzhledem k původnímu dotazu vyhledávané v jiných objemech a se snižující se vyhledávaností klesá jejich relevance. Pro naši metodu budeme vycházet z tzv. rozšířené shody (tedy množiny všech dotazů obsahující daný řetězec). Navrhovaný způsob výpočtu velikosti jednotlivých množin m vychází z předpokladu, že relevance dalších frází z navrhovaných nejhledanějších dotazů je přímo úměrná hledanosti v rozšířené shodě.

Nechť n je zvolená konstanta označující počáteční hloubku analýzy (tedy množství stránek ve startovní množině m_1) a k je tzv. koeficient rozdělení k (tento koeficient říká, jakou váhu budou mít množiny vyšších řádů m oproti m_1). Pak $|m_1| = n$ a $|m_2| + |m_3| + \dots |m_i| = k.n$. Poměry velikostí jednotlivých množin m vyššího řádu odpovídají poměrům volných shod klíčových frází náležícím těmto množinám. Jestliže je velikost některé množiny menší než 1, označí se její velikost za rovnou 0. Množství generovaných množin lze rovněž omezit tzv. zarážkou z, ta označuje řád množiny, jejíž velikost bude automaticky označena jako nulová (např. z=5 znamená, že metoda bude počítat pouze 3 množiny, tedy m_2, m_3 a m_4).

5.3 Ukázka výpočtu velikosti množin m

Předpokládejme použití výše uvedené metody výpočtu velikosti množin h a následující konfiguraci kalibračních konstant:

Počáteční hloubka analýzy $n = 100$

Koeficient rozdělení $k = 1.0$

Zarážka $z = 8$

Dále uvažujme, že jako referenční vyhledávač použijeme seznam.cz a že uživatel zadal slovo „angličtina“. Statistika hledanosti tohoto dotazu je následující:

Rozšířená shoda (dotazy obsahující "angličtina")

	Za den	Za týden
Minimální	645	10 741
Maximální	6 357	36 393
Průměrná	3 938	27 568

Obrázek 5-1 rozšířená shoda dotazu „angličtina“

Přesná shoda (dotazy pouze "angličtina")

	Za den	Za týden
Minimální	67	1 056
Maximální	594	2 825
Průměrná	317	2 220

Obrázek 5-2 přesná shoda dotazu „angličtina“

Nejhledanější dotazy obsahující "angličtina"

	Hledaný výraz	Rozšířená shoda	Přesná shoda
1	angličtina	3 938	317
2	anglictina pro	767	0
3	anglictina online	449	240
4	anglictina zdarma	232	49
5	anglictina pro zacatecniky	221	171
6	anglictina slovesa	212	4
7	anglictina testy	201	119
8	anglictina nepravidelna slovesa	159	146
9	anglictina pro samouky	138	106
10	anglictina nejmensi pro	135	92
11	anglictina deti	134	1

Z tohoto přehledu je patrné, že lidé vyhledávají často pomocí specifičtějšího dotazu než jen „angličtina“, nejčastěji se v rozšířené shodě vyskytuje výraz „angličtina pro“ – tento výraz znázorňuje fakt, že uživatelé hledají angličtinu pro konkrétnější skupinu (začátečníky, samouky, nejmenší, pokročilé, atd.). Těchto skupin může být mnoho a některé se v dalších hledaných výrazech i vyskytují (např. angličtina pro začátečníky). Jelikož naše metoda ignoruje shodné stránky, můžeme klidně použít překrývající se slova. Naopak tímto obdržíme nové stránky a domény budou více odpovídat objemu vyhledávání (výsledky vyhledávání pro slovní spojení „angličtina pro“ a

„angličtina pro začátečníky“ jsou totiž odlišné). Rovněž lze vidět, že je pro tuto metodu velmi důležité oddělení velikosti h1 od způsobu tvorby velikosti množin vyšších řádů. Velikost m1 není záměrně počítána pomocí poměru její rozšířené shody vzhledem k dalším rozšířeným shodám, jelikož by takto byl příliš oslaben vliv dalších nalezených slov. Takto lze vliv dalších domén řídit pomocí koeficientu rozdělení, což je pro naše experimentální účely výhodnější. Ostrá verze této metody by pravděpodobně tento koeficient již nepoužívala, tato modifikace bude probrána v rozšířeních této metody.

Co se samotných výpočtů týče, m1 = n a další množiny závisí na jejich velikostech objemů vyhledávání v rozšířených shodách. Jelikož jsme zvolili z=8, budeme počítat pouze množiny m1..m7. Přiřazení jednotlivých klíčových frází a objemů vyhledávání je následující:

m_1 – angličtina	objem se neuvažuje
m_2 – angličtina pro	767
m_3 – angličtina online	449
m_4 – angličtina zdarma	232
m_5 – angličtina pro začátečníky	221
m_6 – angličtina slovesa	212
m_7 – angličtina testy	201

Nyní víme, že $|m_2| + ... + |m_7| = n.k = 100$. Musíme tedy tyto body rozdělit mezi jednotlivé množiny ve stejném poměru, jako jsou jejich objemy vyhledávání. Výpočtem velikosti jednotlivých množin dle uvedených skutečností, dostáváme následující hodnoty:

$$|m_1| = 100$$
$$|m_2| = 37$$
$$|m_3| = 22$$
$$|m_4| = 11$$
$$|m_5| = 11$$
$$|m_6| = 10$$
$$|m_7| = 9$$

Celkem se tedy bude analyzovat 200 stránek

5.4 Možné modifikace a vylepšení způsobu výpočtu velikosti množiny

Jak již bylo zmíněno, uvedená metoda je určena spíše pro účely experimentování. K tomuto ji předurčuje poměrně dobrá škálovatelnost pomocí výše uvedených konstant. V praxi by se tato metoda pravděpodobně modifikovala. Zde je přehled možných modifikací.

- Počet analyzovaných stránek by se odvozoval z vhodné funkce pro dané objemy vyhledávání. Nebylo by tedy třeba stanovovat konstantu n ani žádné další konstanty. Metoda by se sama přizpůsobovala náročnosti dotazu. Předpokládá se totiž, že více hledané dotazy jsou více zatížené SEO spamem. Metoda by tedy přirozeně věnovala větší pozornost často vyhledávaným dotazům. Díky použití cashe by však tento fakt neměl mít efekt na rychlost samotného vyhledávání. Rovněž určení počtu nutných množin by záviselo na této funkci – funkce by například skončila při dosažení limitního objemu vyhledávání nebo při vytvoření určitého počtu množin. Problémem této modifikace je nalezení této vhodné funkce.
- Filtrování příliš podobných dotazů. Tato modifikace vychází z faktu, že příliš podobné dotazy by mohli zkreslit obsazení množin. Jedná se o efekt, kdy 2 dotazy jsou si tak podobné, že výsledky vyhledávání jsou téměř shodné. Pokud předpokládáme, že navržená metoda přeskakuje totožné stránky (viz 5.1 Princip navrhované metody), můžeme přeskočit příliš mnoho stránek a dostat se na nižší pozice výsledků vyhledávání, které již nemusí být zcela relevantní a přínosné. Proto toto vylepšení předpokládá, že po dosažení určitého počtu shod se bez ohledu na počet nalezených stránek množina uzavře a pokračuje se další množinou.
- Počítání přesné shody místo rozšířené shody. Tato modifikace trochu mění pohled na určování, které dotazy se použijí pro sestavování množin m. Nevychází se zde z podobností dotazů, ale z nejpravděpodobnějšího dalšího konkrétního dotazu.
- Používání součtu přesné a volné shody pro určení objemu vyhledávání.
- Přizpůsobení důležitosti jednotlivých množin dle výkyvů ve vývoji vyhledávání jejich slov. Toto vylepšení by bylo obtížnější na implementaci, nicméně se jedná o zajímavý nástroj, kterým lze přizpůsobit výsledky vyhledávání popularitě toho, co se v danou dobu hledá. Uvažujme případ, kdy jsou vyhlášeny mimořádné volby do poslanecké sněmovny. Pokud bychom vzali dlouhodobé statistiky, budou zatíženy slovy jako volby 2006, komunální volby, atd. navíc se do nich nové slovo „mimořádné volby“ ještě nestihne dostat (jelikož vyhlášení těchto voleb bylo proneseno teprve před několika dny). Pokud bychom ale sledovali velmi prudký nárůst vyhledávanosti klíčového slova „mimořádné volby“ mohli bychom uměle pro toto slovo vytvořit speciální množinu, jejíž stránky by měli větší

ohodnocení rp (například v závislosti na tom, jak velký by byl procentuální přírůstek objemu vyhledávání pro toto slovo). Použití tohoto vylepšení by mohlo do značné míry pomoci vyhledávačům odtušit „co měl uživatel vlastně na mysli“. Dalším příkladem může být například období začátku semestru, kdy lze vypozorovat zvýšení výskytů klíčových frází typu „vysokoškolské koleje“,“koleje VUT“, atd. jelikož studenti hledají informace o kolejích. Pokud tedy někdo zadá v tuto dobu do vyhledávače slovo „koleje“ vyhledávač díky použití těchto speciálních množin správně odhadne, že uživatel pravděpodobně myslel koleje na ubytování nikoliv vlakové a poskytne tedy uživateli kvalitnější výsledky vyhledávání.

Výše uvedené modifikace byly jen nástřelem možností použití této metody. Tyto modifikace lze kombinovat s modifikacemi uvedenými v bodě 5.1 (například použití hubů) a sestavit tak velmi komplexní metodu zkvalitňující výsledky vyhledávání a potlačující efekt SEO spamu.

5.5 Určení míry SEO spamu pomocí metody korelace výsledků vyhledávání

Jak již bylo řečeno hlavním smyslem této metody je odfiltrovat z výsledků vyhledávání stránky, které by své pozice získali díky SEO spamu. Z této metody lze tedy odvodit několik metrik, kterými se dá určit míra SEO spamu použitého na jednotlivé stránky.

Uvažujme množinu stránek m_1, která byla získána aplikací metody korelace výsledků vyhledávání na dostatečně vyhledávaný dotaz. Lze předpokládat, že tento dotaz je atraktivní pro SEO spamery a tedy že některé stránky v uvažované množině budou díky SEO spamu na vysokých pozicích. Uvažujme stránku $s \in m_1$, u které budeme zkoumat míru SEO spamu φ. Pro toto určení můžeme použít některou z následujících metod.

5.5.1 Metoda rozdílu SERPů

Tato metoda vychází z předpokladu, že se SEO spamerům podaří pro stránku s dosáhnout poměrně vysokých pozic v rámci SERP1, ale po aplikovaní metody korelace výsledků vyhledávání bude pozice s v rámci SERP2 podstatně nižší. Důvodem toho propadu je fakt, že na stránku odkazují pouze nerelevantní stránky. Většinou se spamerům podaří získat poměrně slušné výsledky hromadným nakupování zpětných odkazů (stránka získá vysokou pozice v SERP1). Pokud však tyto odkazy jsou z větší části netematické, daná stránka po aplikaci metody bude na tím nižších pozicích v rámci SERP2 čím je větší podíl netematických odkazů směřujících na tuto stránku. Vzorec pro určení φ (míry SEO spamu) je tedy následující:

$\varphi(s) = \frac{SERP1(s) - SERP2(s)}{n} .100 ; n = |SERP|$

Při použití této metody se můžeme setkat s 2mi případy

- Výsledkem je kladné číslo – stránka je tedy ovlivněna SEO spamem. Výsledek udává z kolika procent je pozice ve vyhledávači tímto faktem ovlivněna.
- Výsledkem je záporné číslo – stránka není ovlivněna SEO spamem. Výsledek udává jakou mírou (v procentech) je stránka znevýhodněna v SERPu díky faktu, že jiné stránky používají SEO spam.

5.5.2 Metoda vztahu trp

Tato metoda vychází z předpokladu, že čím vyšší pozice v SERPu stránka dosáhne, tím vyšší by měla být její relevantnost vzhledem k zadanému dotazu (tedy její trp). Lze tedy říci, že každá pozice v rámci tohoto dotazu vyžaduje určitou míru relevance. Pokud stránka tuto míru po aplikaci korelace výsledků vyhledávání nesplňuje, znamená to, že tato stránka používá SEO spam, jehož míra odpovídá podílu relevance, kterou stránka dosáhla a relevance, kterou by stránka musela dosáhnout pro získání původní pozice (v SERP1):

$$\varphi(s) = \frac{trp(s)}{trp(s_x)} ; SERP1(s) = SERP2(s_x)$$

Při použití této metody se můžeme setkat se 2mi případy:

- $\varphi(s) \in (0, 1)$ - stránka používá SEO spam jehož míra odpovídá tomuto číslu
- $\varphi(s) \in < 1, \infty)$ – stránka nepoužívá SEO spam

5.5.3 Metody analýzy více faktorů

Předchozí metody pracují s analýzou výstupu korelovaných a nekorelovaných výsledků vyhledávání. Toto však nemusí být jediné hledisko, kterým by se měla posuzovat míra SEO spamu. Použití těchto metod je opět velmi variabilní a nabízí propojení na další metody – můžeme tedy analyzovat více faktorů:

- analýza rychlosti nárůstu trp je schopna odhalit SEO spam na úrovni cíleného získávání tematických odkazů. Tato modifikace kombinuje metodu, kterou používá společnost google [doplnit citaci na patent] s metodou korelace výsledků vyhledávání.
- analýza anchorů zpětných odkazů při výpočtu trp. Tato modifikace se opět snaží rozpoznat SEO spam na úrovni získávání tematických odkazů. Může se zde analyzovat jak okolí zpětných odkazů, tak jejich variantnost. Velikost přírůstků trp je pak ovlivněna těmito faktory.

- provedení několika iterací metody korelace výsledků vyhledávání (například na množiny m vyšších řádů).
- analýza onpage faktorů (zde je možné například zjišťovat textové okolí zpětných odkazů)
- postih reciprokých zpětných odkazů (tedy kdy 2 stránky na sebe vzájemně odkazují). Toto rozšíření bylo implementováno, jelikož jej považuji za velmi důležité, neboť v českém internetovém prostředí jsou lišty se zpětnými odkazy velmi oblíbené.

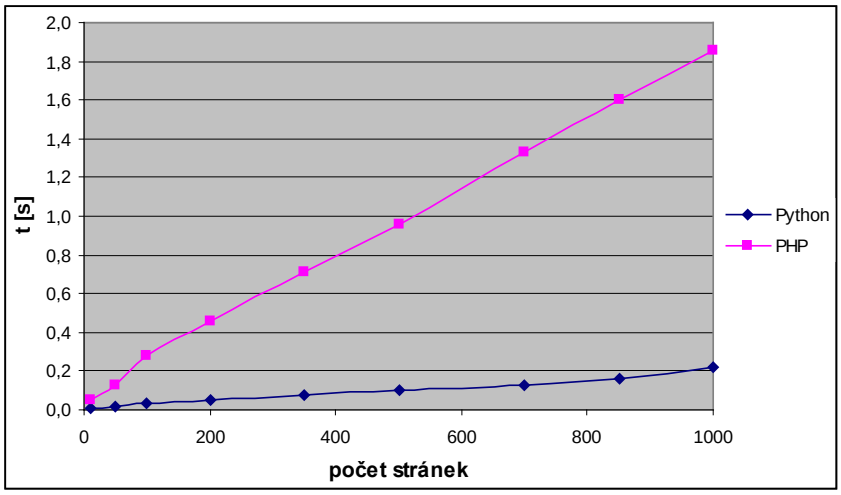
6 Implementace metody korelace výsledků vyhledávání

Tato kapitola popisuje některé implementačně stěžejní aspekty práce. Jedná se o popis realizace uživatelského rozhraní a paralelizace zpracování.

6.1 Implementační prostředí

Z důvodu vyšší přenositelnosti byla metoda implementována v jazyce Python 2.6. Jako vývojové prostředí bylo použito NetBeans IDE 6.7.1. Metoda byla zprovozněna na webovém serveru s podporou jazyků PHP 5.2 a Python 2.6.

Jako alternativa k jazyku Python se nabízelo využití jiného skriptovacího jazyka – například jazyka PHP 5.2. Metoda byla původně v jazyce PHP 5.2 vyvíjena, ale při průběžném testování bylo zjištěno, že časová náročnost při zpracovávání regulárních výrazů je, v poměru s jinými jazyky, neuspokojivá viz následující obrázek:

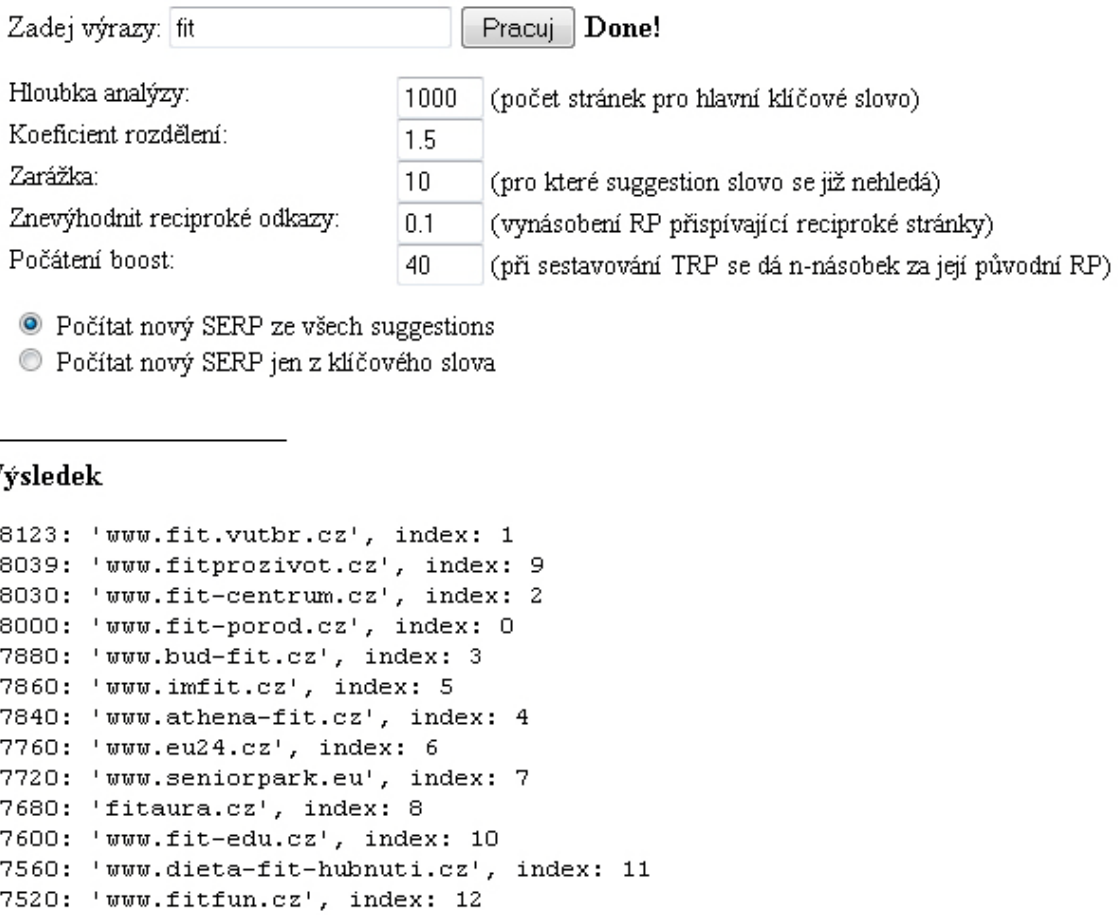


Obrázek 6-1 Graf časové náročnosti zpracování regulárních výrazů

Z tohoto důvodu používá i společnost Seznam.cz, a.s. při vývoji vlastního vyhledávače momo další i jazyk Python [35]. Od PHP varianty bylo tedy ustoupeno. Z této původní implementace se pro finální verzi zachovalo jen uživatelské rozhraní, protože běžně dostupné webové servery s podporou jazyka Python jsou současně schopny interpretovat i jazyk PHP, a jelikož samotná interpretace uživatelského rozhraní nemá na rychlost metody vliv.

6.2 Uživatelské rozhraní

Pro testování implementované metody bylo nutné vytvořit uživatelské rozhraní, prostřednictvím něhož by bylo možné zadávat relevantní parametry pro výpočet. Toto rozhraní bylo implementováno v jazyce PHP 5.2, který ve webovém prohlížeči interpretuje výsledný vzhled pomocí jazyka XHTML a CSS. Rozhraní je složeno z jednoduchého formuláře určeného pro modifikaci zadání parametrů pro metodu a výpisem výsledků tvořené mimo názvu domény i celkovým Relevancy points (v levé části) a původní pozicí ve vyhledávači www.seznam.cz (za slovem index vpravo) – viz následující obrázek zobrazující výsledky vzorového dotazu „fit“:



Obrázek 6-2 uživatelské rozhraní implementované metody

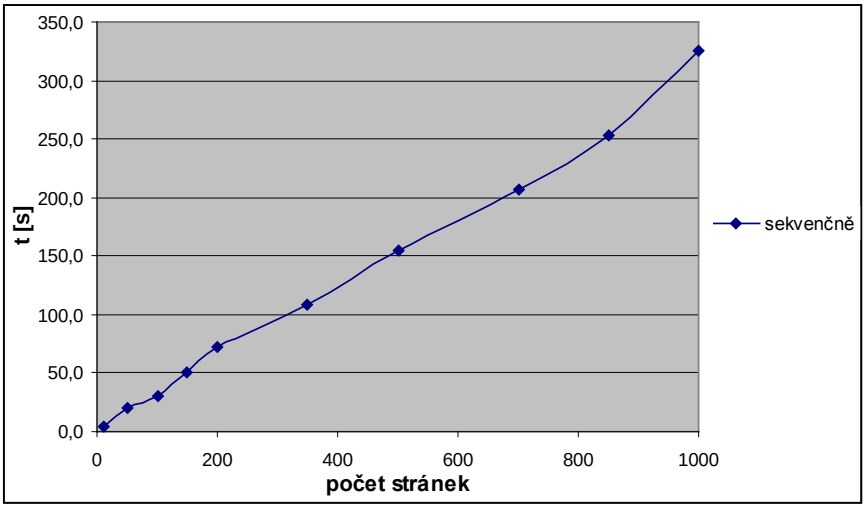
Metoda korelace výsledků vyhledávání, která je napsána v jazyce Python, je volána z uživatelského prostředí pomocí PHP funkce exec() ve tvaru:

$$exec('python\ \{název_souboru\}\ \{hloubka_analýzy\}\ \{koeficient_rozdělení\}\ \{zarážka\}\ \{znevýhodnění_reciprokých_odkazů\}\ \{počáteční_boost\}\ \{počítat_SERP_ze_všech_suggestions\}');$$

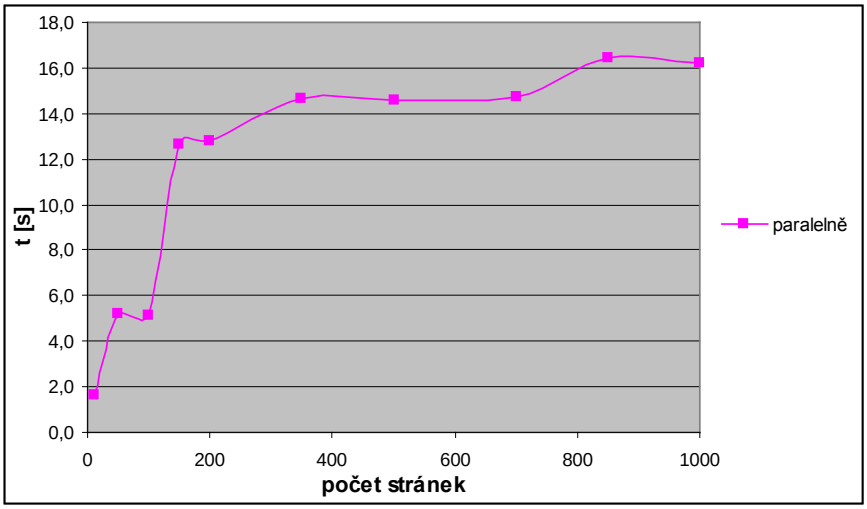
Po dokončení výpočtu metody se již naformátovaný výsledek předá zpět uživatelskému prostředí, kde je následně vypsáno pomocí jazyka html.

6.3 Paralelizace zpracování

Časově nejnáročnějším problémem při výpočtu metody bylo samotné získávání obsahu webových stránek. Tento faktor způsobuje schopnost rychlosti vzdáleného serveru odpovědět na požadavek získání obsahu webu. Muselo se zde tedy přistoupit k paralelizaci downloadu obsahu požadovaných webových stránek. Při sekvenčním zpracování rostl čas se zvyšujícím počtem zkoumaných stránek lineárně (viz Obrázek 6-3). V případě paralelismu byl časový nárůst logaritmický (viz Obrázek 6-4). Toto chování je způsobeno rychlostí odeslání obsahu ze vzdáleného serveru s nejdelší časovou odezvou.



Obrázek 6-3 časy bez paralelizace



Obrázek 6-4 časy s paralelizací

Paralelizace zpracování tedy téměř zdvacetinásobila rychlost implementované metody.

7 Tvorba kalibračních dat

Jak již bylo uvedeno v kapitole 4, k testování úspěšnosti každé nové metody je vhodné použít tzv. kalibrační data. Jelikož jsem během vývoje této metody neměl přístup k žádným kalibračním datům, nemohl jsem těžit ze statisticky významného vzorku (alespoň několik tisíc ohodnocených dotazů).

V rámci testování této metody jsem požádal několik rodinných příslušníků a známých, aby ohodnotili výsledky vyhledávání ze serveru seznam.cz a výsledky vyhledávání poskytnuté implementovanou metodou. Jako vzorek byly náhodně zvoleny následující dotazy:

- Encyklopedie
- Seznam doktorů
- Ploty
- Angličtina
- Brno
- Facebook
- Titulky
- Nábytek
- Kurzy měn
- Zájezdy
- Prezident České republiky

Tyto dotazy byly zvoleny s ohledem na zvýšenou míru vyhledávanosti (v rozdílných úrovních). Lze tedy předpokládat, že tyto dotazy jsou zajímavé pro spamery a tedy některé z jejich výsledků mohou být zatíženy seo-spamem.

Metodika hodnocení výsledů vyhledávání na tyto dotazy byla pomocí ruční klasifikace do jedné z pěti kalibračních kategorií popsaných v kapitole 4. Tato klasifikace byla provedena pro prvních 10 stránek z výsledků vyhledávání na každý kalibrovaný dotaz vždy 5ti uživateli. Každé stránce byl přiřazen počet bodů kvality, který odpovídal průměrnému bodovému ohodnocení podle následující tabulky:

Klasifikovaná kategorie	Počet bodů
Vitální	5
Užitečná	4
Dobrý výsledek	3
Špatný výsledek	2
Off-topic	0

Pro každý dotaz bylo vyhodnoceno prvních 10 stránek ze SERPu vyhledávače seznam.cz a prvních 10 stránek vyhodnocených implementovanou metodou.

7.1 Průměrné ohodnocení jednotlivý stránek výsledků vyhledávání

Seznam průměrných hodnot výsledků vyhledávání získaných ze zdrojů seznam.cz a pomocí implementované metody byl zapsán pro každý dotaz do níže uvedených tabulek. Sloupek *body (B)* obsahuje aritmetický průměr hodnocení, sloupek *vážené body (VB)* obsahuje počet bodů s ohledem na získanou pozici podle následujícího vzorce:

$$VB = B\,.(11 - pozice)$$

Zde jsou výsledky jednotlivých dotazů:

Dotaz	**Vyhledávač**	**Poz**	**Stránka**	**B**	**VB**
Encyklopedie	Seznam	1	cs.wikipedia.org/wiki/Hlavní_strana	5	50
		2	encyklopedie.seznam.cz/	4,9	44,1
		3	www.encyklopedie-knihy.cz/	3,6	28,8
		4	www.cojeco.cz/	4,9	34,3
		5	www.celysvet.cz/	3,4	20,4
		6	www.az-encyklopedie.info/	3,1	15,5
		7	www.mzv.cz/jnp/cz/encyklopedie_statu/index. html	3,1	12,4
		8	encyklopedie.brna.cz/home-mmb/	3	9
		9	www.encyklopedie-vina.cz/	1,1	2,2
		10	www.garten.cz/enc/cz/	2,3	2,3
			celkem získaných bodů	**34,4**	**219**

Encyklopedie	Metoda	1	encyklopedie.seznam.cz	4,9	49
		2	cs.wikipedia.org/wiki/Hlavn%C3%AD_strana	5	45
		3	www.encyklopedie-knihy.cz	3,6	28,8
		4	www.cojeco.cz	4,9	34,3
		5	www.celysvet.cz	3,4	20,4
		6	www.az-encyklopedie.info	3,1	15,5
		7	www.mzv.cz/jnp/cz/encyklopedie_statu/index.html	3,1	12,4
		8	encyklopedie.brna.cz/home-mmb	3	9
		9	www.encyklopedie-vina.cz	1,1	2,2
		10	www.garten.cz/enc/cz	2,3	2,3
			celkem získaných bodů	**34,4**	**218,9**

Poměr úspěšnosti implementované metody vůči vyhledávači seznam.cz pro vyhledávací dotaz Encyklopedie je 1 respektive 0,99 pro vážené body.

Dotaz	**Vyhledávač**	**Poz**	**stránka**	**B**	**VB**
Seznam lékařů	Seznam	1	www.pribram.cz/index.php?menu=seznam-doktoru	3,1	31
		2	cs.wikipedia.org/wiki/Seznam_lékařů	4,4	39,6
		3	opetmuzem.cz/seznam-lekaru.html	2,4	19,2
		4	www.vzp.cz/cms/internet/cz/Pobocky/Jihocesky_kr	3,5	24,5
		5	www.cck-br.cz/seznam-lekaru/	3,4	20,4
		6	www.racerbuggy.cz/index.php/seznam-lekaru	2,1	10,5
		7	www.boiron.cz/search_region.php?d=CZ0&lg=1&	4	16
		8	www.budtezdravi.cz/lekari.php	3,7	11,1
		9	www.duchcov.cz/vismo/dokumenty2.asp?id_org=3	2,9	5,8
		10	www.allivictus.cz/?nav=cz/poradna/seznam-lekaru/	0,4	0,4
			celkem získaných bodů	**29,9**	**178,5**
Seznam lékařů	Metoda	1	www.ordinace.com	4	40
		2	www.pribram.cz/index.php?menu=seznam-doktoru	3,1	27,9
		3	http://www.pracovni-	3,2	25,6
		4	cs.wikipedia.org/wiki/Seznam_lékařů	4,4	30,8
		5	www.vzp.cz/cms/internet/cz/Pobocky/Jihocesky_kr	3,5	21
		6	opetmuzem.cz/seznam-lekaru.html	2,4	12
		7	www.cck-br.cz/seznam-lekaru/	3,4	13,6
		8	www.racerbuggy.cz/index.php/seznam-lekaru	2,1	6,3
		9	www.boiron.cz/search_region.php?d=CZ0&lg=1&	4	8
		10	www.budtezdravi.cz/lekari.php	3,7	3,7
			celkem získaných bodů	**33,8**	**188,9**

Poměr úspěšnosti implementované metody vůči vyhledávači seznam.cz pro vyhledávací dotaz seznam lékařů je 1,13 respektive 1,06 pro vážené body.

dotaz	**vyhledávač**	**Poz**	**Stránka**	**B**	**VB**
Ploty	seznam	1	www.ploty-pletivo.cz	4	40
		2	www.e-pletivo.cz	4	36
		3	www.dops.cz	4	32
		4	www.apleg-ploty.cz	3,6	25,2
		5	www.ploty-betafence.cz	4,3	25,8
		6	www.meziploty.cz/cs/	0,2	1
		7	www.plotovecentrum.cz/	3,5	14
		8	www.ploty-saka.cz	3,1	9,3
		9	www.vrata-lamark.cz	2,8	5,6
		10	www.ploty-pletivo-oploceni.cz	4	4

			celkem získaných bodů	**33,5**	**192,9**
Ploty	metoda	1	www.e-pletivo.cz	4	40
		2	www.ploty-pletivo.cz	4	36
		3	www.ploty-betafence.cz	4,3	34,4
		4	www.dops.cz	4	28
		5	www.plotyabrany.cz	4	24
		6	www.apleg-ploty.cz	3,6	18
		7	www.dirickx.cz	3,8	15,2
		8	www.meziploty.cz/cs/	0,2	0,6
		9	www.plotovecentrum.cz/	3,5	7
		10	www.vrata-lamark.cz	2,8	2,8
			celkem získaných bodů	**34,2**	**206**

Poměr úspěšnosti implementované metody vůči vyhledávači seznam.cz pro vyhledávací dotaz Ploty je 1,02 respektive 1,07 pro vážené body.

Dotaz	**Vyhledávač**	**Poz**	**Stránka**	**B**	**VB**
Angličtina	seznam	1	www.helpforenglish.cz/	4,8	48
		2	www.anglictina-on-line.cz/	4	36
		3	www.aktivnistudium.cz/anglictina.php	3,1	24,8
		4	www.anglictinacz.com/	3,5	24,5
		5	www.anglictina-hrou.cz/	3,2	19,2
		6	cs.wikipedia.org/wiki/Angličtina	2,6	13
		7	www.anglictina-hry.cz/	2,5	10
		8	www.meriva-translations.com/	1,1	3,3
		9	www.english-online.cz/	4,4	8,8
		10	www.anglictina-v-anglii.cz/	1,6	1,6
			celkem získaných bodů	**30,8**	**189,2**
Angličtina	metoda	1	www.english-online.cz	4,4	44
		2	www.dynamicenglish.cz	2,2	19,8
		3	www.anglictina-hry.cz/	2,5	20
		4	www.helpforenglish.cz/	4,8	33,6
		5	www.aktivnistudium.cz/anglictina.php	3,1	18,6
		6	www.anglictina-on-line.cz/	4	20
		7	www.anglictinacz.com/	3,5	14
		8	cs.wikipedia.org/wiki/Angli%C4%8Dtina	3,4	10,2
		9	www.anglictina-hrou.cz/	3,2	6,4
		10	www.aj.cz	3,3	3,3
			celkem získaných bodů	**34,4**	**189,9**

Poměr úspěšnosti implementované metody vůči vyhledávači seznam.cz pro vyhledávací dotaz Angličtina je 1,12 respektive 1,00 pro vážené body.

Dotaz	**Vyhledávač**	**Poz**	**Stránka**	**B**	**VB**
Brno	seznam	1	www.brno.cz/	5	50
		2	hc-kometa.cz	2	18
		3	www.info-brno.cz/index.php	4	32
		4	www.1fcbrno.cz/index.php	1,8	12,6
		5	www.fnbrno.cz	3,2	19,2
		6	www.bvv.cz/ibf	3,2	16
		7	www.dpmb.cz/	3,1	12,4
		8	www.stred.brno.cz/	3,6	10,8
		9	brno.idnes.cz/	4,1	8,2

		10	www.airport-brno.cz	2,6	2,6
			celkem získaných bodů	**32,6**	181,8
Brno	metoda	1	www.brno.cz	5	50
		2	www.dpmb.cz	3,1	27,9
		3	hc-kometa.cz	2	16
		4	www.info-brno.cz/index.php	4	28
		5	www.fnbrno.cz	3,2	19,2
		6	www.1fcbrno.cz	1,8	9
		7	www.ibrno.cz	4,6	18,4
		8	www.bvv.cz/ibf	3,2	9,6
		9	brno.idnes.cz	4,1	8,2
		10	www.stred.brno.cz	3,6	3,6
			celkem získaných bodů	**34,6**	189,9

Poměr úspěšnosti implementované metody vůči vyhledávači seznam.cz pro vyhledávací dotaz Brno je 1,06 respektive 1,04 pro vážené body.

Dotaz	**Vyhledávač**	**poz**	**Stránka**	**B**	**VB**
Facebook	seznam	1	www.facebook.com/?_fb_noscript=1	5	50
		2	cs-cz.facebook.com/?_fb_noscript=1	0	0
		3	www.meebo.com/	4	32
		4	www.dobryweb.cz/skoleni-socialni-site-twitter-	2,1	14,7
		5	www.facemag.cz/	4	24
		6	www.facelook.cz/	1,1	5,5
		7	m.facebook.com/	0,2	0,8
		8	cs.wikipedia.org/wiki/Facebook	3,5	10,5
		9	www.lupa.cz/clanky/proc-nepouzivam-facebook/	1,4	2,8
		10	www.pooh.cz/a.asp?a=2016048	3,4	3,4
			celkem získaných bodů	**24,7**	**143,7**
Facebook	metoda	1	www.facemag.cz	4	40
		2	www.facebook.com/?_fb_noscript=1	5	45
		3	www.dobryweb.cz/skoleni-socialni-site-twitter-	2,1	16,8
		4	cs-cz.facebook.com/?_fb_noscript=1	0	0
		5	cs.wikipedia.org/wiki/Facebook	3,5	21
		6	www.meebo.com	4	20
		7	www.com-facebook.cz	3,1	12,4
		8	www.facelook.cz	1,1	3,3
		9	m.facebook.com	0,2	0,4
		10	www.lupa.cz/clanky/proc-nepouzivam-facebook	1,4	1,4
			celkem získaných bodů	**24,4**	**160,3**

Poměr úspěšnosti implementované metody vůči vyhledávači seznam.cz pro vyhledávací dotaz Facebook je 0,99 respektive 1,12 pro vážené body.

Dotaz	**Vyhledávač**	**poz**	**Stránka**	**B**	**VB**
titulky	seznam	1	www.titulky.com	5	50
		2	www.opensubtitles.org/cs	4,9	44,1
		3	http://titulky.gamepark.cz/	4,8	38,4
		4	http://mojetitulky.com/	4,3	30,1
		5	http://www.serialzone.cz/titulky/	4	24
		6	subs.cz	3,5	17,5
		7	titulky-	4	16
		8	www.sga-project.com/sgtt/	1,6	4,8

		9	www.mysubtitles.com/titulky-cesky	2,5	5
		10	subtitles.sg1.cz	1,6	1,6
			celkem získaných bodů	**36,2**	**231,5**
titulky	metoda	1	www.titulky.com	5	50
		2	http://mojetitulky.com/	4,3	38,7
		3	www.opensubtitles.org/cs	4,9	39,2
		4	http://titulky.gamepark.cz/	4,8	33,6
		5	http://www.serialzone.cz/titulky/	4	24
		6	titulky-	4	20
		7	subs.cz	3,5	14
		8	subtitles.sg1.cz	1,6	4,8
		9	www.sga-project.com/sgtt/	1,6	3,2
		10	www.mysubtitles.com/titulky-cesky	2,5	2,5
			celkem získaných bodů	**36,2**	**230**

Poměr úspěšnosti implementované metody vůči vyhledávači seznam.cz pro vyhledávací dotaz titulky je 1,00 respektive 0,99 pro vážené body.

Dotaz	**Vyhledávač**	**poz**	**Stránka**	**B**	**VB**
nábytek	seznam	1	www.nabytek-forliving.cz	3,5	35
		2	www.sconto.cz	4,3	38,7
		3	www.nabytek-eno.cz	3,1	24,8
		4	www.jena-nabytek.cz	3,5	24,5
		5	www.spectrumnabytek.cz	3,4	20,4
		6	www.dreveny-nabytek.cz	2,8	14
		7	www.asko-nabytek.cz	4,2	16,8
		8	www.nabytek-homemarket.cz	3,5	10,5
		9	www.nejlevnejsinabytek.cz	3,5	7
		10	www.mona-nabytek.cz	1,6	1,6
			celkem získaných bodů	**33,4**	**193,3**
nábytek	metoda	1	www.nabytek-forliving.cz	3,5	35
		2	www.nabytek-homemarket.cz	3,5	31,5
		3	www.sconto.cz	4,3	34,4
		4	www.jena-nabytek.cz	3,5	24,5
		5	www.spectrumnabytek.cz	3,4	20,4
		6	www.nabytek-eno.cz	3	15
		7	www.interier.com	4,3	17,2
		8	www.dreveny-nabytek.cz	2,8	8,4
		9	www.asko-nabytek.cz	4,2	8,4
		10	www.nejlevnejsinabytek.cz	3,5	3,5
			celkem získaných bodů	**36**	**198,3**

Poměr úspěšnosti implementované metody vůči vyhledávači seznam.cz pro vyhledávací dotaz nábytek je 1,08 respektive 1,03 pro vážené body.

Dotaz	**Vyhledávač**	**poz**	**Stránka**	**B**	**VB**
Kurzy měn	seznam	1	www.kurzy.cz	4,8	48
		2	www.patria.cz	4,5	40,5
		3	www.finance.cz	4	32
		4	kurzy.pozri.cz	3,2	22,4
		5	www.akcie.cz	3,8	22,8
		6	www.kurzy-men-online.cz	3,8	19
		7	www.atlantik.cz/zpravodajstvi/kurzy-men	3,3	13,2

		8	velkejkuk.cz/devizove-kurzy-cnb/meny	2,5	7,5
		9	www.online-kurzy-men.cz	3	6
		10	kurzy-men.aloki.cz	2,5	2,5
			celkem získaných bodů	**35,4**	**213,9**
Kurzy měn	metoda	1	www.kurzy.cz	5	50
		2	kurzy.pozri.cz	3,2	28,8
		3	www.prodejvalut.cz	4	32
		4	www.atlantik.cz	3,3	23,1
		5	www.kurzy-men-online.cz	3,8	22,8
		6	www.akcie.cz	3,8	19
		7	www.finance.cz	4	16
		8	kurzy-men.aloki.cz	2,5	7,5
		9	www.patria.cz/zpravodajstvi/home.html	4,5	9
		10	kurzymen.eu	3,5	3,5
			celkem získaných bodů	**37,6**	**211,7**

Poměr úspěšnosti implementované metody vůči vyhledávači seznam.cz pro vyhledávací dotaz nábytek je 1,08 respektive 1,03 pro vážené body.

Dotaz	**Vyhledávač**	**poz**	**Stránka**	**B**	**VB**
Zájezdy	seznam	1	www.zajezdy.cz/	4,4	44
		2	www.nettravel.cz/	4,3	38,7
		3	www.last-minute.cz/	4,4	35,2
		4	www.firotour.cz/	3,9	27,3
		5	www.invia.cz/	4,5	27
		6	www.cedok.cz/	4,6	23
		7	www.poznavaci-zajezdy.cz/	3,1	12,4
		8	www.etravel.cz	3,7	11,1
		9	airstop.cz/last-minute/	3,6	7,2
		10	last-minute.invia.cz/	0	0
			celkem získaných bodů	**36,5**	**225,9**
Zájezdy	metoda	1	www.last-minute.cz	4,4	44
		2	www.invia.cz	4,5	40,5
		3	www.nettravel.cz	4,3	34,4
		4	www.zajezdy.cz	4,4	30,8
		5	www.firotour.cz	3,9	23,4
		6	www.cedok.cz	4,6	23
		7	www.poznavaci-zajezdy.cz/	3,1	12,4
		8	www.etravel.cz	3,7	11,1
		9	www.ckmayer.cz	3,7	7,4
		10	www.zajezdy.org	4,1	4,1
			celkem získaných bodů	**40,7**	**231,1**

Poměr úspěšnosti implementované metody vůči vyhledávači seznam.cz pro vyhledávací dotaz zájezdy je 1,12 respektive 1,02 pro vážené body.

Dotaz	**Vyhledávač**	**poz**	**Stránka**	**B**	**VB**
prezident	seznam	1	www.novinky.cz/zahranicni	1,1	11
čeké republiky		2	zpravy.idnes.cz/obama-medvedev-a-klaus-cesky-	3,2	28,8
		3	cs.wikipedia.org/wiki/V%C3%A1clav_Klaus	5	40
		4	cs.electionsmeter.com/hlasov%C3%A1n%C3%AD	4,1	28,7
		5	www.mediafax.cz/politika/3039160-Prezident-	2,8	16,8
		6	www.rozhlas.cz/zpravy/evropskezalezitosti	2,5	12,5

		7	www.nwoo.sk/2009/09/cesky-prezident-schuze-	2,7	10,8
		8	www.ct24.cz/volba-prezidenta/cesky-prezident	2,8	8,4
		9	www.blisty.cz/2009/10/14/art49411.html	2,7	5,4
		10	www.radio.cz/cz/clanek/125677	1,8	1,8
			celkem získaných bodů	**28,7**	**164,2**
prezident	metoda	1	zpravy.idnes.cz/obama-medvedev-a-klaus-cesky-	3,2	32
čeké republiky		2	cs.wikipedia.org/wiki/V%C3%A1clav_Klaus	5	45
		3	cs.electionsmeter.com/hlasov%C3%A1n%C3%AD	4,1	32,8
		4	www.novinky.cz/domaci/195198-novy-cesky-	0,9	6,3
		5	www.rozhlas.cz/zpravy/evropskezalezitosti	2,5	15
		6	www.nwoo.sk/2009/09/cesky-prezident-schuze-	2,7	13,5
		7	www.ct24.cz/volba-prezidenta/cesky-prezident	2,8	11,2
		8	www.blisty.cz/2009/10/14/art49411.html	2,7	8,1
		9	www.mediafax.cz/politika/3039160-Prezident-	2,8	5,6
		10	www.radio.cz/cz/clanek/125677	1,8	1,8
			celkem získaných bodů	**28,5**	**171,3**

Poměr úspěšnosti implementované metody vůči vyhledávači seznam.cz pro vyhledávací dotaz prezident české republiky je 0,99 respektive 1,04 pro vážené body.

8 Testování úspěšnosti implementované metody

Výše uvedené výsledky vyhledávání byly sestaveny dne 12.5.2010 a pro srovnání byla implementovaná metoda nastavena na následující koeficienty:

Hloubka analýzy: **1000** (počet stránek pro hlavní klíčové slovo)

Koeficient rozdělení: **1,5** (poměr pro rozdělení mezi množinou m_1 a dalšími množinami)

Zarážka: **10** (pro které našeptávané slovo se již nehledá)

Znevýhodnit reciproké odkazy: **0,1** (vynásobení RP přispívající reciproké stránky)

Počátení boost: **40** (při sestavování TRP se dá n-násobek za její původní RP)

Jelikož jsem neměl dostatek kalibračních dat pro nalezení ideálních koeficientů, byly tyto koeficienty zvoleny s použitím následujících úvah:

- Hloubka analýzy by měla být co největší, aby metoda byla co nejpřesnější (zde bohužel nemohu procházet více než 100 stránek, jelikož výchozí vyhledávač (seznam.cz) neposkytuje více výsledků vyhledávání
- Koeficient rozdělení 1,5 znamená, že se bude brát v úvahu 100% RP z počáteční množiny m_1(generované původním hledaným výrazem) a 50% RP z množin generovaných našeptávanými slovy. Tímto rozdělením jsem omezil váhu našeptaných slov a tím snížil

riziko kritických chyb, kdy by se do výsledků vyhledávání dostali zcela nerelevantní stránky.

- Zarážku jsem zvolil maximální možnou – tedy 10, jelikož seznam.cz uveřejňuje maximálně 10 příbuzných výrazů.
- Konstantu pro znevýhodnění reciprokých odkazů jsem nastavil na číslo 0,1 z důvodu, že jsem se snažil co nejvíce postihnout vzájemnou výměnu odkazů mezi stránkami (takto byl rapidně snížen vliv „seo lišt“).
- Počáteční boost 40 jsem nastavil proto, aby se respektovala počáteční síla stránky vzhledem k velké hloubce analýzy.

8.1 Srovnání výsledků, úspěšnost metody

Z uvedených tabulek v kapitole 7 vyplívá, že metoda měla různé úspěšnosti pro různé dotazy. Bohužel z důvodů nedostatku kalibračních dat nebylo možné rozpoznat, zda je metoda úspěšnější pro korelaci velmi konkurenčních slov (ve smyslu SEO optimalizátorů) jako jsou například „zájezdy“ nebo specifičtějších slov (např. „seznam lékařů“). Všeobecně lze říci, že tato metoda by neměla být použita sama o sobě, ale pouze jako doplněk dalších metod (proto byl v této metodě zaveden počáteční boost). Implementovaná metoda se však ukázala jako velmi užitečná, neboť v některých případech dostala do SERPu uživateli velmi kladně hodnocené stránky, které v původních prvních 10ti výsledcích vyhledávání poskytnutých vyhledávačem seznam.cz chyběly. Jednalo se především o stránky www.ibrno.cz, www.interier.com, www.prodejvalut.cz, www.ordinace.com, www.zajezdy.org. Naopak se podařilo z první strany výsledků vyhledávání odstranit uživateli velmi špatně hodnocené stránky jako například www.mona-nabytek.cz, www.anglictina-v-anglii.cz/, www.allivictus.cz/?nav=cz/poradna/seznam-lekaru/ a další. Z naměřených výsledků lze shrnout následující statistické údaje:

- Celkový počet bodů získaných vyhledávačem seznam.cz: **356,1**
- Celkový počet bodů získaných pomocí implementované metody: **374,8**
- Celkový počet vážených bodů získaných vyhledávačem seznam.cz: **2133,9**
- Celkový počet vážených bodů získaných pomocí implementované metody: **2196,3**

Implementovaná Metoda korelace výsledků vyhledávání tedy **zlepšuje bodové ohodnocení** uživatelů na uvedeném vzorku dotazů vyhledávač seznam.cz průměrně o **5%** pokud neuvažujeme umístění výsledků v rámci první stránky nebo též o **3%** pokud zohledňujeme i umístění (vážené body).

9 Závěr

Hlavním cílem této práce bylo především prozkoumat technologie internetových vyhledávačů a metody boje se seo-spamem. Práce do značné míry narážela na problém, že reálně využívané metody samotné vyhledávače tají a nelze tedy s určitostí říci, jak moc je která metoda úspěšná. I přes to je zde popsána většina principů, na kterých pravděpodobně staví i komerční vyhledávače. Studium těchto principů významně přispělo k návrhu vlastní autorské metody. Tato metoda bohužel nemůže využít rozsáhlé databáze stránek a jejich parametrům, kterými komerční vyhledávače disponují, nicméně je schopna ze vzorků v řádu tisíce stránek sestavit výsledky vyhledávání, které do určité míry odbourávají seo-spam. Tato metoda byla rovněž testována na vzorku stránek ohodnocených pěti uživateli (tzv. kalibrační metodou) a dosahovala výsledků téměř o 3% lepších, než poskytnul komerčně nejúspěšnější vyhledávač v České Republice - seznam.cz. Nejenom vývoj implementované metody ale i studium technologií používaných vyhledávači je neustálý proces, jelikož seo-spameři každý den vyvíjí nové techniky, kterými lze zmanipulovat výsledky vyhledávání. Proto si myslím, že metoda navržená v této diplomové práci by měla být pravidelně rozšiřována a testována. Věřím, že tato práce bude přínosem či inspirací nejen pro všechny vývojáře vyhledávačů, ale rovněž pro všechny zájemce o SEO problematiku.

Literatura

[1] R. HENZINGER , Monika , MOTWANI, Rajeev , SILVERSTEIN, Craig . Challenges in web search engines. *ACM SIGIR Forum*. 2002, vol. 36, is. 2, s. 11-22. Dostupný z WWW: <http://portal.acm.org/citation.cfm?id=792550.792553>.

[2] HARRITY & SNYDER, LLP. United States Patent Application: 0050071741 : Information retrieval based on historical data [online]. 2005 [cit. 2009-11-14]. Dostupný z WWW: <http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.html&r=1&f=G&l=50&s1=%2220050071741%22.PGNR.&OS=DN/20050071741&RS=DN/20050071741>.

[3] *United States Patent: 7617200 : Displaying context-sensitive ranked search results* [online]. 2007 [cit. 2009-11-13]. Dostupný z WWW: <http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PTXT&s1=google&s2=search.TI.&OS=google+AND+TTL/search&RS=google+AND+TTL/search>.

[4] NEMRAVA, Jan. *Optimalizace WWW stránek pro vyhledávací a indexovací katalogy*. [s.l.], 2004. 75 s. Vysoká škola ekonomická v Praze Fakulta managementu v Jindřichově Hradci Katedra managementu informací. Vedoucí diplomové práce Ing. Jiří Jelínek, CSc.

[5] TODO doplnit zdroj „reality“

[6] Google. *Google's Search Engine Optimization Starter Guide*. [s.l.] : [s.n.], 2008. 22 s. Dostupný z WWW: <http://googlewebmastercentral.blogspot.com/2008/11/googles-seo-starter-guide.html>.

[7] Seznam a.s.. *Výměna odkazů* [online]. 2009 [cit. 2009-12-18]. Dostupný z WWW: <http://napoveda.seznam.cz/cz/fulltext-hledani-v-internetu/vymena-odkazu/>.

[8] Google. *Pokyny pro webmastery* [online]. 2009 [cit. 2009-12-18]. Dostupný z WWW: <http://www.google.com/support/webmasters/bin/answer.py?answer=35769>.

[9] Google. *Skrytý text a odkazy* [online]. 2009 [cit. 2009-12-19]. Dostupný z WWW: <http://www.google.com/support/webmasters/bin/answer.py?answer=66353>.

[10] Google. *Kanonizace* [online]. 2009 [cit. 2009-12-20]. Dostupný z WWW: <http://www.google.com/support/webmasters/bin/answer.py?answer=139066>.

[11] *United States Patent Application: 0050071741* [online]. 2009 [cit. 2009-12-21]. Dostupný z WWW: <http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.html&r=1&f=G&l=50&s1=%2220050071741%22.PGNR.&OS=DN/20050071741&RS=DN/20050071741>.

[12] Google. *Paid links* [online]. 2009 [cit. 2009-12-21]. Dostupný z WWW: <http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=66736>.

[13] List of Google PageRank updates 2010 [online]. 2010 [cit. 2010-02-15]. Dostupný z WWW: <http://www.1websitedesigner.com/google-pagerank#dates>.

[14] JANOVSKÝ, Dušan. Jak Psát Web [online]. 2010 [cit. 2010-02-14]. Google PageRank. Dostupné z WWW: <http://www.jakpsatweb.cz/seo/pagerank.html>.

[15] BRIN, Sergey , PAGE, Lawrence . The Anatomy of a Large-Scale Hypertextual Web Search Engine [online]. 1998 [cit. 2010-02-15]. Dostupný z WWW: <http://infolab.stanford.edu/~backrub/google.html>.

[16] PageRank Distribution Removed From WMT - Webmaster Central Help [online]. 2010 [cit. 2010-02-15]. Dostupný z WWW: <http://www.google.com/support/forum/p/Webmasters/thread?tid=6a1d6250e26e9e48&hl=en>.

[17] D. Fogaras and B. Rácz. Towards scaling fully personalized PageRank. In Proceedings of the 3rdWorkshop on Algorithms and Models for theWeb-Graph (WAW), strany 105–117, Řím, Itálie, 2004.

[18] BENCZÚR, András A., et al. SpamRank – Fully Automatic Link Spam Detection : Work in progress. [s.l.] : [s.n.], 2009. 14 s.

[19] Google ltd. Statistiky vyhledávání Google [online]. 2010 [cit. 2010-02-12]. Dostupný z WWW: <http://www.google.com/insights/search/#q=motory&date=today%201-m&cmpt=q>.

[20] HAVELIWALA, Taher H. TopicSensitive PageRank. Stanford, 2002. 10 s. Oborová práce. Stanford University Computer Science Department Stanford, CA 94305. Dostupné z WWW: <http://citeseer.ist.psu.edu/haveliwala02topicsensitive.html>

[21] KLEINBERG, JON M. Authoritative Sources in a Hyperlinked Environment. New York, 1999. 32 s. Oborová práce. Cornell University, Ithaca, New York.

[22] GYÖNGYI, Zoltan ; GARCIA-MOLINA, Hector ; PEDERSEN, Jan . Combating Web Spam with TrustRank. Stanford : Computer Science Department, 2004. 12 s.

[23] RAJ, Rashmi. Web Spam Detection with AntiTrust Rank. Stanford, CA 4305, 2006. 6 s. Oborová práce. Stanford University Computer Science Department Stanford, CA 94305. Vedoucí práce Vijay Krishnan. Dostupné z WWW: <http://www.google.com/url?sa=t&source=web&ct=res&cd=1&ved=0CAYQFjAA&url=http%3A%2F%2Fciteseerx.ist.psu.edu%2Fviewdoc%2Fdownload%3Fdoi%3D10.1.1.92.481%26rep%3Drep1%26type%3Dpdf&ei=SXrAS6XcOobEOLW06JYE&usg=AFQjCNGjcf6S7YNBEGcjMN7iUq6HjsVg8Q&sig2=rKVt1kryVC2mc8uB6RGCKg>.

[24] JANOVSKÝ, Dušan. BadRank, seznamte se. Yuhůův weblog o webu [online]. 2008, 10, [cit. 2010-02-11]. Dostupný z WWW: <http://weblog.jakpsatweb.cz/d/1225209900-badrank-seznamte-se.html>.

[25] BECCHETTI, Luca . Using Rank Propagation and Probabilistic Counting for LinkBased Spam Detection [online]. Rome, Italy : Yahoo! Research Barcelona, Spain & Santiago, Chile, 2006. 10 s. Oborová práce. DIS Universit `a di Roma “La Sapienza”. Dostupné z WWW: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.6845>.

[26] LUCA, Becchetti, et al. LinkBased Characterization and Detection of Web Spam. Barcelona, Spain & Santiago, Chile : Yahoo! Research, 2006. 8 s.

[27] Microsoft Research Strider Team, in collaboration with UCD. Microsoft research [online]. 2006 [cit. 2010-03-19]. Strider Search Defender: Automatic and Systematic Discovery of Search Spammers through Non-Content Analysis. Dostupné z WWW: <http://research.microsoft.com/en-us/um/redmond/projects/strider/searchdefender/>

[28] Microsoft research [online]. 2006, 2010 [cit. 2010-03-19]. Strider URL Tracer with Typo-Patrol. Dostupné z WWW: <http://research.microsoft.com/en-us/um/redmond/projects/strider/urltracer/>.

[29] SLAWSKI, Bill . SEO by the Sea [online]. 2006 [cit. 2010-04-01]. Phrase Based Information Retrieval and Spam Detection. Dostupné z WWW: <http://www.seobythesea.com/?p=413>.

[30] Domeček z karet [online]. 1. července 2009 [cit. 2010-04-28]. Jak fungují automatické odkazové lišty. Dostupné z WWW: <http://www.domecekzkaret.cz/jak-funguji-automaticke-odkazove-listy>.

[31] PASSADORE, Andrea ; GROSSO, Alberto ; BOCCALATTE, Antonio . An Agent-based Semantic Search Engine for Scalable Enterprise Applications. University of Genova, DIST, Via Opera Pia 13, 16145 Genova, Italy : University of Genova, 2009. 13 s. Dostupné z WWW: <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-460/paper08.pdf>.

[32] KEYHANIPOOR, Amir Hossein , et al. A Multi-Layer/Multi-Agent Architecture for Meta-Search Engines. AIML 05 Conference : Control and Intelligent Processing Center of Excellence, Electrical and Computer Eng. Department, University of Tehran , 2005. 4 s. Dostupné z WWW: <http://www.icgst.com/AIML05/papers/P1120535110.pdf>.

[33] JASEN, James. First [online]. 2010 [cit. 2010-05-09]. Using an Intelligent Agent to Enhance Search Engine Performance. Dostupné z WWW: <http://131.193.153.231/www/issues/issue2_3/jansen/index.html>.

[34] TOPlist [online]. 2010 [cit. 2010-05-09]. Globální statistika. Dostupné z WWW: <http://www.toplist.cz/global.html>.

[35] Seznam.cz. In Wikipedia : the free encyclopedia [online]. St. Petersburg (Florida) : Wikipedia Foundation, 20.8.2006, last modified on 6.2.2009 [cit. 2010-05-19]. Dostupné z WWW: <http://cs.wikipedia.org/wiki/Seznam.cz>.

Seznam příloh

Příloha 1. CD – zdrojové soubory

Seznam obrázků

Obrázek 2-1 statistika dotazu z vyhledávače seznam.cz 10
Obrázek 3-1 odkazování hubů na autoritativní stránky 19
Obrázek 3-2 závislost mezi huby a autoritami 22
Obrázek 3-3 ukázka pravděpodobného spamu 23
Obrázek 3-4 šíření anti-trust ranku 25
Obrázek 3-5 počet spamových stránek podle použitých metod s různými parametry 25
Obrázek 3-6 ukázka linkové farmy 26
Obrázek 3-7 graf distribuce odkazů 27
Obrázek 3-8 funkce damping faktoru 27
Obrázek 3-9 použití truncated PageRanku s různou konstantou T 28
Obrázek 3-10 výchozí chování spamerů podle výzkumného týmu Strider 29
Obrázek 3-11 *Demonstrace „ovládnutí“ googlu. Pořízeno 9.5.2010* 33
Obrázek 3-12 nasazení agentů ve vyhledávačích nové generace 34
Obrázek 4-1 podíl vyhledávačů na českém trhu 35
Obrázek 5-1 rozšířená shoda dotazu „angličtina“ 41
Obrázek 5-2 přesná shoda dotazu „angličtina“ 41
Obrázek 6-1 Graf časové náročnosti zpracování regulárních výrazů 46
Obrázek 6-2 uživatelské rozhraní implementované metody 47
Obrázek 6-3 časy bez paralelizace 48
Obrázek 6-4 časy s paralelizací 49