

Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky

Katedra informačních technologií

Studijní program : Aplikovaná informatika

Obor : Informační systémy a technologie

Data-driven SEO

DIPLOMOVÁ PRÁCE

Student : Bc. Jiří Koutný

Vedoucí : Ing. Martin Žamberský

Oponent : Ing. Petr Kováčík

2012

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracoval samostatně a že jsem uvedl všechny použité prameny a literaturu, ze které jsem čerpal.

V Praze 1. 12. 2012

.....

Bc. Jiří Koutný

Poděkování

Děkuji vedoucímu své práce, Ing. Martinu Žamberskému, za trpělivost a všechny připomínky, kterými mi pomohl práci zkvalitnit.

Abstrakt

Obor optimalizace pro vyhledávače (SEO) v poslední době prochází velkými změnami. Na trhu se konečně objevily nástroje, které umožňují marketingovým konzultantům výsledky své práce v SEO efektivně měřit a vyhodnocovat.

Teoretická část této práce si proto klade za cíl popsat a porovnat vybrané zajímavé SEO nástroje včetně konkrétních možností jejich praktického využití. Hlavní prostor je věnován databázím zpětných odkazů (MajesticSEO, SEOMoz OpenSiteExplorer a Ahrefs) a nástrojům pro návrh klíčových slov od Googlu (AdWords), Seznamu (Sklik) a okrajově také Wordtracker a SEMRush. Závěr teoretické části obsahuje přehled nástrojů a postupů pro měření pozic ve vyhledávačích.

Praktická část práce popisuje způsob výběru, přípravy a zpracování dat ze zmíněných nástrojů pro korelační analýzu výsledků vyhledávače Seznam.cz ve vztahu k nejznámějším SEO faktorům. Výsledky analýzy pomůžou marketingovým konzultantům ujasnit si, které faktory jsou pro úspěch ve vyhledávačích nejdůležitější a má smysl věnovat úsilí jejich zlepšování.

Klíčová slova

seo, linkbuilding, klíčová slova, data-driven seo, marketing ve vyhledávačích, optimalizace pro vyhledávače, korelační analýza.

Abstract

The Search Engine Optimization (SEO) industry has recently undergone major changes. Many new analytics tools have been put on the market enabling marketing consultants to be finally able to measure and evaluate the results of their work in SEO effectively.

The **theoretical part** of this diploma thesis therefore aims to describe and compare selected SEO tools including practical examples of their use. The paper is focused on backlink databases (MajesticSEO, SEOMoz OpenSiteExplorer and Ahrefs) and keyword suggestion tools from Google (AdWords), Seznam (Sklik), Wordtracker and SEMRush. The final chapter provides an overview of search engine positions tracking tools and techniques.

The **practical part** describes the method of selection, preparation and processing of data obtained from tools mentioned above. The data are used to compute correlation analysis of Seznam.cz search engine results in relation with the best known SEO factors. The results of the analysis will help marketing consultants to clarify which factors are the most important to focus on to obtain more traffic from search engines.

Keywords

seo, linkbuilding, keywords, data-driven seo, search marketing, search engine optimization, correlation analysis.

Obsah

1. Úvod	1
1.1. Vymezení tématu a důvod jeho výběru.....	1
1.2. Cíle práce.....	2
1.3. Přínosy práce a vlastní přínos	3
1.4. Předpoklady práce	3
1.5. Omezení práce	3
1.6. Cílová skupina práce	4
1.7. Rešerše literatury	4
1.7.1. Česká literatura.....	4
1.7.2. Zahraniční literatura.....	5
2. Vymezení SEO	6
2.1. Co je cílem SEO?	7
2.2. Co není cílem SEO?	8
2.3. Co je data-driven SEO?.....	9
2.3.1. Seznam data-driven SEO nástrojů použitých v práci.....	10
2.3.2. Příklady stanovení priorit v SEO na základě dat	10
2.3.3. Kvalitativní vyhodnocování prováděných změn	11
3. Nástroje a techniky data-driven SEO	12
3.1. Databáze zpětných odkazů	12
3.1.1. Vlastnosti odkazů.....	12
3.1.2. Popis a porovnání vybraných databází	15
3.1.3. Vlastní porovnání velikosti indexů odkazových databází.....	17
3.1.4. Kvalitativní srovnání nástrojů	20
3.1.5. Cenové srovnání vybraných nástrojů	22
3.1.6. Reálné kalkulace využívání API.....	24
3.1.7. Kterou odkazovou databází zvolit?	25
3.1.8. Praktické využití: analýza konkurence.....	27
3.2. Nástroje pro analýzu klíčových slov.....	29
3.2.1. Popis vybraných nástrojů pro analýzu klíčových slov	29

3.2.2. Vlastní porovnání.....	31
3.2.3. Praktické využití: návrh struktury webu.....	32
3.2.4. Praktické využití: Porovnání značek.....	37
3.3. Nástroje pro měření pozic ve vyhledávačích.....	39
3.3.1. Jak se pozice ve vyhledávačích měří?.....	40
3.3.2. Jak se scraping provádí?.....	41
3.3.3. Problémy se stránkováním.....	42
3.3.4. Co je potřeba brát v úvahu při získávání výsledků vyhledávačů.....	44
3.3.5. Právní aspekty scrapingu.....	44
3.3.6. Detekce problémů a nečekaných situací na stráně vyhledávače.....	46
4. Korelační analýza výsledků vyhledávače Seznam.cz.....	48
4.1. Cíle.....	48
4.2. Výběr atributů (SEO faktorů) korelace.....	48
4.3. Vybrané atributy korelace.....	49
4.3.1. Titulek stránky.....	50
4.3.2. Nadpisy stránky (H1 – Hx).....	50
4.3.3. Atributy ostatního obsahu stránky.....	51
4.3.4. Atributy domény a URL.....	52
4.4. Hypotézy.....	53
4.5. Výběr klíčových slov pro analýzu.....	54
4.6. Jak Collabim získává data o pozicích webů ve vyhledávačích?.....	56
4.7. Zpracování výsledků.....	57
4.7.1. Stažení obsahu stránek, které se objevily ve výsledcích hledání Seznam.cz.....	57
4.7.2. Výpočet hodnot všech SEO faktorů pro jednotlivé stránky.....	58
4.7.3. Výpočet korelačního koeficientu pro jednotlivá klíčová slova.....	60
4.7.4. Korelační koeficienty pro jednotlivé SEO faktory.....	64
4.8. Vyhodnocení hypotéz.....	66
4.9. Kompletní porovnání výsledků s odhady důležitosti serveru SEOfactory.cz.....	67
4.10. Domény, které se nejčastěji objevovaly v analyzovaných výsledcích.....	68
4.11. Hlavní problémy, které jsem při zpracování práce musel řešit.....	69
5. Závěr.....	71

5.1. Naplnění cílů.....	71
5.2. Možnosti rozšíření práce	71
6. Terminologický slovník.....	73
7. Použité zdroje	74
8. Seznam obrázků	77
9. Seznam tabulek	79
10. Přílohy	80
Příloha 1: Schéma databáze použité pro korelační analýzu.....	80
Příloha 2: Předložky a spojky, které byly při normalizaci odstraněny z textu analyzované stránky	80
Příloha 3: Algoritmus pro výpočet pořadového ranku vstupů Spearmanovy korelace	81

1. Úvod

1.1. Vymezení tématu a důvod jeho výběru

V oboru optimalizace pro vyhledávače (SEO) sleduji v poslední době velký posun od SEO řízeného názory a domněnkami k SEO, které je řízené tvrdými daty a kvantitativními analýzami. Pro druhou skupinu se začíná používat termín **data-driven SEO**, viz např. video Randa Fishina ze SEOMoz (1). Jiní autoři používají také termín analytics-driven SEO (2).

Na českém internetu je bohužel o data-driven SEO, jeho technikách a potřebných nástrojích stále nedostatek kvalitních článků nebo publikací. Důsledkem je situace, kdy mnoho internetových podnikatelů, ale i „profesionálních“ SEO konzultantů stále měří SEO pouze na základě dosažených pozic svých webů ve vyhledávačích.

U právníků se s nadsázkou říká, že „dva právníci, tři různé názory na problém“. U SEO konzultantů je situace téměř stejná. I v roce 2012 jsou česká diskusní fóra o SEO plná „zaručených“ rad sebevědomých SEO konzultantů, jak lépe optimalizovat web.

Stačí si přečíst několik příspěvků: *„Titulek stránky má přeci pro vyhledávač vyšší váhu než text příchozích odkazů.“* Další uživatel samozřejmě striktně nesouhlasí a své tvrzení dokládá vlastním (ač nesmyslným) průzkumem: *„Nesouhlasím. Titulek jsem optimalizoval na všech třech svých webech a žádný efekt to nemělo. Stačilo ale získat 5 odkazů s textem ‚levné batohy‘ a hned mi pozice vystřelily vzhůru!“* Tyto rady nejsou ničím jiným, než špatně podloženými domněnkami.

Podobné debaty se bohužel vedou již od příchodu SEO na český Internet v roce 2001, kdy Marek Prokop ve svém článku *„Optimalizace stránek pro vyhledávací a indexovací služby“* (3), pravděpodobně jako první použil český termín **„Optimalizace pro vyhledávače“**.

V této práci chci podobným debatám alespoň částečně udělat přítrž a pomoci českým SEO konzultantům dělat to, co je v SEO skutečně důležité a co jejich klientům vydělá více peněz.

Vybrané téma mě zajímá také z důvodu, že jsem dva roky pracoval jako SEO konzultant ve firmě H1.cz, která je přední internetovou agenturou v České republice. Po svém odchodu z H1.cz již téměř tři roky provozuji český webový SEO nástroj Collabim¹, který patří mezi nejznámější na českém trhu. Nejen s daty z nástroje Collabim budu pracovat v této práci.

1.2. Cíle práce

Hlavními cíli práce bylo:

1. popsat a porovnat vybrané analytické nástroje využitelné pro data-driven SEO,
2. vybrat a připravit data pro korelační analýzu SEO faktorů pro vyhledávač Seznam.cz,
3. realizovat korelační analýzu na základě získaných dat, vyhodnotit výsledky.

Pro účel splnění všech hlavních cílů jsem si definoval také **dílčí cíle**, které jsou nezbytné pro úspěšné splnění cílů hlavních:

1. definovat základní pojmy SEO a zařadit tyto termíny v rámci celého internetového marketingu,
2. popsat co je a také není cílem SEO,
3. definovat pojem data-driven SEO a osvětlit základní principy fungování nástrojů pro data-driven SEO,
4. porovnat popisované nástroje mezi sebou,
5. uvést praktické příklady využití popisovaných nástrojů,
6. vybrat vhodné SEO faktory pro korelační analýzu a definovat hypotézy pro důležitost těchto faktorů,
7. naprogramovat aplikaci, která získá a připraví vstupní data pro korelační analýzu,

¹ <http://www.collabim.cz>

8. vyhodnotit závěry vzešlé z korelační analýzy.

1.3. Přínosy práce a vlastní přínos

Práce popisuje a porovnává nástroje, které česká odborná veřejnost buď ještě vůbec nezná, nebo zná pouze jejich základní možnosti využití. Srovnání pomůže SEO konzultantům uvědomit si přednosti i slabé stránky jednotlivých nástrojů a podle toho si vybrat ty, které jim budou nejvíce užitečné.

Korelační analýza pomůže všem internetovým marketérům vytvořit si lepší přehled o **důležitosti jednotlivých SEO faktorů u vyhledávače Seznam.cz.**

1.4. Předpoklady práce

Předpokladem splnění definovaných cílů práce byl především **přístup k popisovaným SEO nástrojům a jejich datům.** Mnoho nástrojů je k dispozici zdarma. U placených nástrojů jsem většinou využil bezplatného zkušebního období. K nástroji Collabim a jeho datům o pozicích webů na vyhledávači Seznam.cz jsem měl, jako jeho majitel, neomezený přístup.

Práci bych také nebyl schopen vytvořit bez zkušeností, které mi dala **dvouletá práce ve firmě H1.cz**, téměř tříleté provozování SEO nástroje Collabim a **aktivní zájem o obor SEO.** Tyto zkušenosti mi umožnily definovat reálné problémy, které SEO konzultanti a jejich klienti řeší, a ve kterých by jim popisované SEO nástroje mohly být nápomocné.

Neméně důležitým předpokladem byla pokročilá **znalost programování a nastavování serveru**, bez které bych nemohl vytvořit algoritmus získání a zpracování dat z vyhledávače Seznam.cz.

1.5. Omezení práce

Teoretická část práce si neklade za cíl popsat všechny SEO nástroje pro data-driven SEO. Popisuje a srovnává především ty, jejichž data jsou použita v praktické části práce.

Při korelační analýze jsem SEO faktory sledoval na aktuálním obsahu analyzovaných stránek. Je samozřejmě nutné počítat s faktem, že vyhledávače mohou mít ve svém indexu mnohem **starší verzi daných stránek se zcela odlišnými parametry** (např. zcela rozdílené texty nebo odkazy).

1.6. Cílová skupina práce

Práce přinese cenné poznatky všem SEO konzultantům a internetovým marketérům, kteří chtějí dělat svá rozhodnutí na základě dat z kvalitních analytických nástrojů místo dojmů a pocitů.

1.7. Rešerše literatury

1.7.1. Česká literatura

Česká literatura o SEO je typická zaměřením na začínajícího uživatele. Pravděpodobně z důvodu malého trhu/cílové skupiny nevznikla na českém trhu žádná kniha o SEO zaměřená na středně pokročilé až pokročilé čtenáře.

Z české oborové literatury o SEO je zajímavý především titul *Velký průvodce SEO* od Michala Kubíčka (4). Tato kniha je ale dnes, téměř 2 roky po svém vydání, již dost zastaralá. Především kapitoly popisující chování vyhledávačů Seznam.cz, Centrum.cz nebo Google jsou velice neaktuální a pro čtenáře mohou být matoucí.

V současnosti mám rozečtenou novou knihu *50 způsobů, jak získat zpětný odkaz* také od Michala Kubíčka (5). V knize jsou velmi jednoduše zodpovězeny základní otázky linkbuildingu a autor se snaží vyvracet některé zažité nesmysly z tohoto oboru.

Aktuálnější a odborněji zaměřené jsou některé bakalářské a diplomové práce. Způsoby analýzy zpětných odkazů shrnuje konzultantka agentury H1.cz Linda Hlaváčová v práci *Analýza odkazového profilu* (6). Autorka pravděpodobně jako první na českém Internetu popisuje do hloubky možnosti odkazové databáze MajesticSEO, kterou ve své práci také zmiňuje.

V SEO stejně jako v celém internetovém marketingu je nutné pro udržení přehledu číst především oborové blogy, které suplují neexistující knihy pro pokročilejší čtenáře.

Doslova nadčasové jsou články od Marka Prokopa (Sova v síti²). V poslední době velmi dobře píší také Lukáš Pítra³, Martin Matějka⁴ nebo Martin Kolčaba⁵.

² <http://www.sovavsiti.cz/>

³ <http://www.lukaspitra.cz/blog-2/>

1.7.2. Zahraniční literatura

V zahraničí je k dispozici mnohem více kvalitních autorů než v Čechách. Zmíním např. druhé vydání knihy *The Art of SEO* (7) od skupiny autorů v čele se CEO známého serveru SEOMoz.org, Randem Fishkinem.

Kniha podrobně popisuje všechny části SEO od návrhu architektury webu, přes analýzu klíčových slov, tvorbu obsahu až po získávání zpětných odkazů pomocí standardních metod i linkbaitingu. Druhé vydání je rozšířeno především o možnosti SEO v prostředí sociálních sítí (Facebook, Twitter).

V zahraničí fungují desítky internetových magazínů zaměřených pouze na SEO. Osobně čas od času čtu velmi populární SEOMoz.org nebo Search Engine Land⁶.

Praktická část mé práce se zaměřuje na vyhledávač Seznam.cz, který funguje pouze v Česku. Z toho důvodu je pro mě česká literatura relevantnější než literatura zahraniční.

⁴ <http://www.martinmatejka.cz/clanky/>

⁵ <http://blog.komart.cz/recenze-linkbuilding/>

⁶ <http://searchengineland.com/>

2. Vymezení SEO

Pro termín SEO (Search Engine Optimization) existuje bezpočet definic. Např.:

„Search Engine Optimization (SEO) – optimalizace stránky pro vyhledávače. Zjednodušeně se jedná o konkrétní techniky zabývající se způsobem, jak umístit na co nejlepší pozici ve vyhledávání.“⁽⁴⁾

„SEO je postup vylepšování a propagace webové stránky, s jehož pomocí stránka zvýší svou návštěvnost z vyhledávačů. SEO má mnoho součástí od textu vašich stránek po způsob, jakým na vaši stránku odkazují ostatní stránky. SEO v některých případech znamená pouze úpravu struktury vašeho webu tak, aby jí vyhledávače rozuměly.“⁷

Mému pohledu na SEO se nejvíce přibližuje definice: *„SEO je proces zvyšování objemu a kvality návštěvnosti z přirozených výsledků hledání vyhledávačů na vybraný web.“* (9)

Nejdůležitějšími složkami SEO jsou dle mého názoru (řazeno sestupně podle důležitosti):

1. vyhledávání a analýza klíčových slov,
2. získávání zpětných odkazů z tematicky příbuzných webů (linkbuilding),
3. psaní webových textů (copywriting),
4. analýza trhu,
5. technická analýza webu.

SEO v rámci internetového marketingového mixu zařazujeme následovně:

➔ Online marketing / Internetový marketing

⁷ „SEO is the practice of improving and promoting a web site in order to increase the number of visitors the site receives from search engines. There are many aspects to SEO, from the words on your page to the way other sites link to you on the web. Sometimes SEO is simply a matter of making sure your site is structured in a way that search engines understand.“⁽⁸⁾

- ➔ Search marketing / Search Engine Marketing (česky Marketing ve vyhledávačích)
- ➔ **SEO** (návštěvnost z **přirozených** výsledků hledání)
 - ➔ Link building
- ➔ PPC reklama (návštěvnost z **placené reklamy** ve výsledcích hledání)

Do stejné skupiny lze k SEO zařadit ještě PPC reklamu, která ale funguje na zcela odlišných principech.

Považuji za důležité zmínit, že aktivity SEO v žádném případě nejsou striktně ohraničeny a často se prolínají s dalšími činnostmi tvorby webu. Jako příklad uvedu návrh informační architektury, copywriting, public relations, user interface/user experience design a webová analytika.

2.1. Co je cílem SEO?

Hlavním cílem SEO je vydělat více peněz. Marketéři by se neměli nechat zmást růstem relevantní návštěvnosti, ba dokonce ani růstem obrátu. Důležité je především tvrdě sledovat náklady a výnosy. (10)

Nárůstu zisku musí předcházet zvýšení návštěvnosti z vyhledávačů optimalizovaného webu. Tyto návštěvníky v případě SEO přivádíme bez nutnosti vyhledávači explicitně platit (tj. kupovat si PPC reklamu ve výsledcích hledání). Mimo kvantitativního aspektu návštěvnosti nás pochopitelně zajímá i její kvalita. Příchodem návštěvníka na náš web totiž vše teprve začíná. Náš web jej musí přesvědčit, aby splnil požadované cíle vedoucí k růstu zisku (zakoupení nabízeného zboží, odeslání poptávky pomocí kontaktního formuláře atd.).

Splnění cíle webu návštěvníkem odborně nazýváme konverze (např. konverze návštěvníka na zákazníka). Zlepšování konverzních schopností webu není přímo obsahem SEO. Jedná se ale o spojené nádoby. Pokud pomocí SEO přivádíme na web návštěvníky, kteří nekonvertují, měli bychom naše postupy zcela přehodnotit.

Se zvýšením návštěvnosti souvisí také zlepšení dohledatelnosti webu ve vyhledávačích. Optimalizovaný web by se měl zobrazovat ve vyhledávači v ideálním případě v prvních

deseti výsledcích na maximální počet tematicky relevantních klíčových slov. Mimo tematické relevance nás samozřejmě zajímají také konverzní schopnosti klíčových slov, na která máme ve vyhledávacích dobré pozice. Pokud například má náš web výborné pozice na všechny modely praček elektrolux, ale tuto značku neprodává, z přivedených návštěvníků u nás pravděpodobně nakoupí jen zlomek.

2.2. Co není cílem SEO?

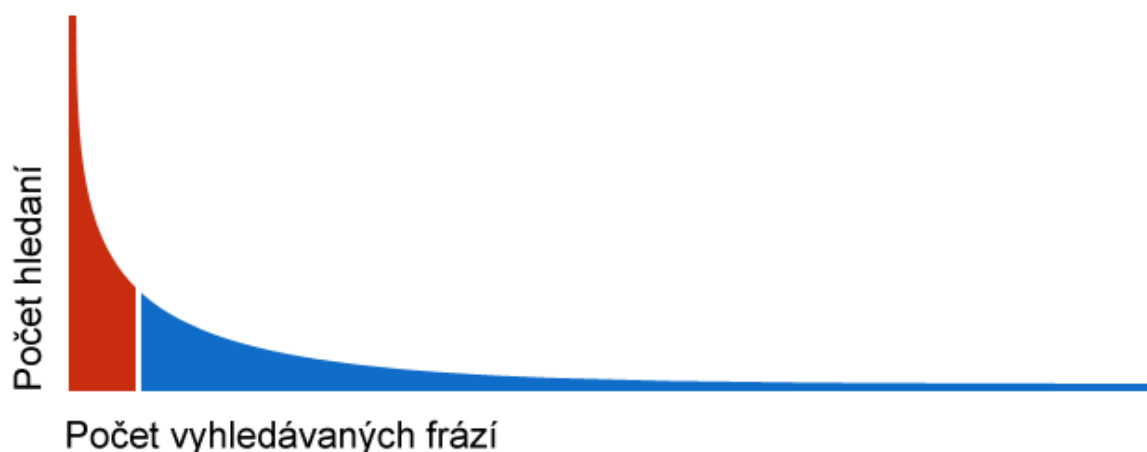
Cílem SEO není získat první pozici ve výsledcích hledání na všechny myslitelné fráze. Vždy bychom se měli soustředit na fráze, které tematicky úzce souvisejí s naším webem. V opačném případě na náš web budou chodit nerelevantní návštěvníci, kteří budou například přetěžovat naše servery a zvyšovat tak naše náklady bez zvýšení příjmů.

(Ne)důležitost prvních pozic výstižně shrnul Dušan Janovský: „*Můžu mít pozici na nějaké slovo, které nikdo nehledá. Nebo můj odkaz na první pozici vypadá tak pitomě, že na něj nikdo neklikne. Nebo mám první pozici na vyhledávači, přes který skoro nikdo nehledá. Na co pak ta pozice je?*“ (11)

Není také vhodné soustředit se pouze na několik vybraných obecných frází jako např. letenky, dovolená, pračky atd. Tato klíčová slova jsou velmi konkurenční a v dnešní době může získání první pozice ve vyhledávači na tato slova stát i stovky tisíc korun.

Obecná slova samozřejmě není vhodné ignorovat a z obsahu webu zcela vypustit. Hlavní kategorie zboží e-shopu se musí pojmenovat obecným slovem (např. pračky). Mimo ně je ale vhodné soustředit se více na slova z tzv. dlouhého ocasu (long tail).

„*Long tail znamená málo vyhledávané (většinou víceslovné) fráze, které však v součtu převyšují hledanost (obecných) klíčových slov. Díky tomu může stránka dosahovat vysoké návštěvnosti, i když se na často hledaná obecná slova (např. nábytek nebo dovolená) nevyskytuje ve výsledcích vyhledávání na předních pozicích. Stačí, aby web obsadil přední místa na desítky podrobnějších frází (např. dřevěný zahradní nábytek)*“ (12)



Obrázek 1: Červená část grafu obsahuje malý počet velmi často hledaných klíčových slov. Naproti tomu modrá část obsahuje velký počet málo hledaných slov (long-tail) (12)

2.3. Co je data-driven SEO?

Data-driven SEO představuje evoluci klasického SEO. První SEO konzultanti měli k dispozici jen velmi málo dat, na základě kterých mohli vymýšlet strategii a cíle optimalizace. Neměli k dispozici pokročilé analytické nástroje, kterými mohli vyhodnocovat úspěšnost optimalizace. SEO tak bylo z velké části o dojmech a pocitech. SEO konzultanti si nemohli být nikdy jisti, co v optimalizaci funguje a co nikoliv. Jen velmi málo z nich bylo schopno svá tvrzení doložit konkrétními měřeními a pozorováními.

S příchodem data-driven SEO se může celý obor optimalizace mnohem rychleji rozvíjet. Autoři konečně mohou používat předchozí měřitelné a důvěryhodné výsledky zkoumání jako základ svých dalších analýz.

V současnosti na trhu existují desítky až stovky placených i bezplatných nástrojů, které kvantitativně zkoumají jednotlivé atributy důležité pro úspěšnou optimalizaci. Máme nástroje pro vyhledávání zpětných odkazů, dokážeme analyzovat textový obsah webů, víme, jak dobře je vyhledávače indexují a mnoho dalšího.

V následujících podkapitolách ukážu z **vlastní zkušenosti** konkrétní příklady, jak je vhodné využívat data pro lepší výsledky optimalizace.

2.3.1. Seznam data-driven SEO nástrojů použitých v práci

Tabulka 1: Seznam data-driven SEO nástrojů použitých v práci. (zdroj: autor)

Název	URL	Kategorie
MajesticSEO	majesticseo.com	databáze zpětných odkazů
SEOMoz Open Site Explorer	opensiteexplorer.org	databáze zpětných odkazů
Ahrefs	ahrefs.com	databáze zpětných odkazů
Google Webmaster Tools	google.com/webmasters/tools	analýza technických faktorů webu
Google AdWords Keyword Tool	adwords.google.cz/o/KeywordTool	nástroj pro návrh klíčových slov
Nástroj pro návrh klíčových slov Skliku	sklik.cz/navrh-klicovych-slov	nástroj pro návrh klíčových slov
Wordtracker	wordtracker.com/find-the-best-keywords	nástroj pro návrh klíčových slov
SEMRush	semrush.com	nástroj pro návrh klíčových slov, databáze zpětných odkazů

Ne všechny nástroje pro data-driven SEO jsou vhodné pro analýzu českých webů. Je to dáno především tím, že u nás má zásadní tržní podíl vyhledávač Seznam.cz, který zahraniční nástroje většinou neznají a nepodporují.

Konkrétní možnosti využití SEO nástrojů budou popsány v dalších kapitolách.

2.3.2. Příklady stanovení priorit v SEO na základě dat

Za pomoci dostupných analytických nástrojů známe podrobné údaje o svém webu i o webech konkurence. Např.:

- Pomocí Google Webmaster Tools jsme zjistili, že míra prokliku našich výsledků z Google je velice nízká. Návštěvnost našeho webu z vyhledávání je proto nedostatečná, přestože stránky našeho webu obsazují ve vyhledávacích přední

příčky.

Vyplývající priorita: upravit texty důležitých stránek webu pro zlepšení návštěvnosti z již dosažených pozic v Google.

- Dle nástroje MajesticSEO má naše konkurence mnohem více zpětných odkazů a odkazuje na ně celkově mnohem více unikátních tematických webů.

Vyplývající priorita: Získávat odkazy z podobných tematických webů jako konkurence nákupem a vymyslet další taktiky, jak je v počtu odkazů předstihnout.

- Analýzou pozic našeho webu ve vyhledávacích zjistíme, že weby přeprdejce našeho zboží mají ve většině případů lepší pozice na klíčová slova obsahující názvy našeho klíčového zboží.

Vyplývající priorita: Zlepšit pozice našeho webu na klíčová slova obsahující názvy našeho zboží a zvýšit tak obrat z našeho klíčového zboží.

2.3.3. Kvalitativní vyhodnocování prováděných změn

U každé změny/experimentu je možné definovat hypotézu a následně ji ověřit na základě reálných výsledků. Příklad:

- **Současný stav:** Naše stránka se zobrazuje na průměrné 3. pozici v Google na klíčové slovo *letenky new york*. Návštěvnost našeho webu z této fráze se nám ale zdá nedostatečná. Podle nástroje Google Webmaster Tools víme, že míra prokliku našeho výsledku z výsledků hledání Google na náš web je 0,5%.
- **Hypotéza:** Zkrácením a přepsáním titulku stránky a přidáním poutavějšího textu do značky *meta description* můžeme zvýšit míru prokliku našeho výsledku a tedy i návštěvnost našeho webu z Googlu na tuto frázi alespoň o 50%
- **Způsob ověření platnosti hypotézy:** Změnu míry prokliku v čase můžeme sledovat pomocí nástroje Google Webmaster Tools. Změnu reálné návštěvnosti z vyhledávače Google určíme například pomocí nástroje Google Analytics v sekci *Traffic sources -> Sources -> Search -> Organic -> filtr Keyword exactly matching „letenky new york“*

3. Nástroje a techniky data-driven SEO

V následujících podkapitolách popíši různé skupiny nástrojů pro data-driven SEO včetně konkrétních příkladů jejich praktického použití. Seznam nástrojů není zdaleka konečný. Kapitola popisuje především nástroje, které jsou následně **využity v praktické části práce**.

Srovnání nástrojů proběhlo mezi **ledem a únorem 2012**. Vybíral jsem do něj nástroje, které jsou v SEO komunitě používány a často citované. Mimo ně samozřejmě existují desítky až stovky nástrojů, které buď nemají tak dobrou kvalitu nebo mají špatný marketing a SEO odborníci včetně mě o nich nevědí.

3.1. Databáze zpětných odkazů

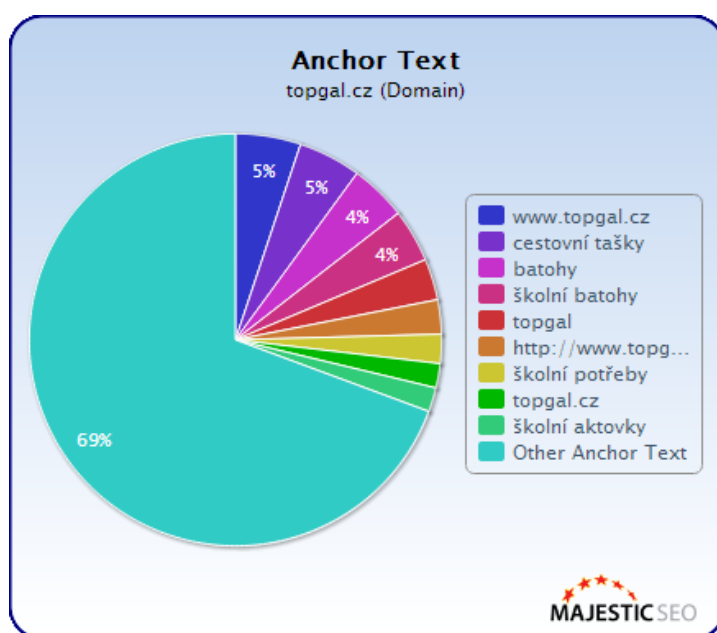
Zpětné odkazy představují jeden z hlavních faktorů, podle kterých vyhledávače hodnotí webové stránky. Právě z toho důvodu **analyzujeme, jak s nimi vyhledávače pracují a jaké faktory u odkazů hodnotí**. Vyhledávače z pochopitelných důvodů neodhalují, které odkazy a do hodnocení stránek započítávají. Z toho důvodu vzniklo několik nástrojů, popsaných níže.

3.1.1. Vlastnosti odkazů

Každý (zpětný) odkaz z pohledu vyhledávače přenáší určitý **rank** (síla odkazu) ze stránky A na stránku B. Základní vlastnosti odkazu (13), slide 7:

- **URL zdrojové (odkazující) a cílové (odkazovaná) stránky odkazu,**
Z pohledu odkazů vyhledávače rozlišují konkrétní stránky webu, ne pouze web jako celek. Největší počet odkazů má v naprosté většině případů titulní stránka webu. V ideálním případě by ale dostatek odkazů měl vést i na konkrétní podstránky. Ty se pak ve výsledcích vyhledávání zobrazují na konkrétnější klíčová slova (např. *digitální zrcadlovky s objektivem*).
- **text odkazu/atribut *alt* u obrázkových odkazů,**
Čím více odkazů s konkrétním klíčovým slovem na náš web vede, tím lepší pozice

ve vyhledávacích náš web na tato klíčová slova získá.



Obrázek 2: Rozložení textů odkazů (anchor text) v odkazech webu topgal.cz dle MajesticSEO. (zdroj: autor)

- **stáří odkazu.**

Vyhledávače sledují, jak se mění odchozí odkazy ze stránky. Pokud se odkazy mění příliš často, může to být pro vyhledávač signál, že se nejedná o přirozeně vytvořený odkaz.

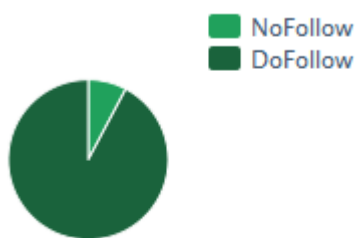
Níže popisované odkazové databáze umožňují sledovat ještě další atributy odkazů, které vyhledávače pravděpodobně **započítávají do svého hodnocení stránek**. Jedná se o:

- **přítomnost atributu rel=nofollow odkazů,**

Google ústy šéfa svého webspam teamu radí webmasterům, aby atributem *rel=nofollow* označovali všechny odkazy, „za které neručí“ (14). Tímto atributem jsou proto většinou označovány odkazy z komentářů, sociálních sítí atd. Pokud vyhledávací robot Googlu narazí na takový odkaz, dává mu pravděpodobně nižší váhu, než odkazům bez atributu *rel=nofollow*.

Z toho důvodu odkazové databáze umožňují tyto odkazy filtrovat, zobrazují jejich

poměr ke všem odkazům webu atd.



Obrázek 3: Poměr dofollow a nofollow odkazů webu Mall.cz dle služby Ahrefs. (zdroj: autor)

- **typy odkazů,**

Odkazy nejsou pouze textové. Mohou mít formu obrázku. IFrame nebo formulář jsou v podstatě také odkazy. Některé služby detekují také odkazy webu, které procházejí přes jedno nebo více přesměrování (např. odkazy z Facebooku mají formu www.facebook.com/l.php?u=http://www.mall.cz, ze které se přesměrují na finální <http://www.mall.cz>)

- **TLD odkazující stránky (.com, .cz, .gov atd.),**

Někteří webmasteri stále věří, že např. odkazům z domén s koncovou .gov dává Google vyšší váhu, než odkazům z ostatních domén.

- **počet stránek webu, který na nás odkazuje,**

Počet stránek odkazujícího webu pomáhá webmasterovi udělat si rychlou představu, jestli je daný web velký a zavedený portál, nebo se jedná o malý bezvýznamný blog.

- **IP adresy webů, které na nás odkazují,**

Jedna populární teorie říká, že odkazy z webů umístěných na různých IP adresách (přesněji IP adresách s různou C-class) mají vyšší váhu, než odkazy z webů IP adresách se stejnou C-class. Odkazové databáze proto umějí odkazy seskupovat právě podle tohoto atributu.

- **počet unikátních domén, z jejichž stránek na nás vedou odkazy,**

Podobně jako počet unikátních IP adres vyhledávače velmi pravděpodobně sledují též počet unikátních domén druhého řádu, ze kterých na konkrétní web vedou odkazy.

- **počet odkazů (=sílu) stránky, která odkazuje na nás,**
Kvalita a kvantita odkazů stránky určují její sílu. Čím více kvalitních odkazů stránka má, tím větší sílu přivede i na náš web v případě, že z ní povede odkaz na náš web.
- **rychlost nárůstu odkazů v čase.**
Příliš rychlý nárůst odkazů webu může vyhledávači signalizovat, že odkazy pro daný web jsou vytvářeny uměle (např. nákupem na odkazových burzách). Tomu se vyhledávače pochopitelně brání a snaží se takové odkazy odfiltrovat nebo celý web různě penalizovat.

3.1.2. Popis a porovnání vybraných databází

Analyzované nástroje: [MajesticSEO](#), [OpenSiteExplorer](#), [Ahrefs](#)

Srovnání vychází z podobného srovnání, uveřejněného na webu RankAbove.com (15).

Uvedené nástroje mají k dispozici vlastní databázi zpětných odkazů a vztahů mezi nimi (který web odkazuje na který). Stejně jako Google nebo Yahoo indexují miliardy webových stránek. Na rozdíl od fulltextových vyhledávačů ale neumožňují prohledávat celý obsah stránek, pouze vztahy mezi nimi (odkazy).

Zmíněné odkazové databáze nabízejí především nástroj **typu Site Explorer**, který zobrazuje metriky zvoleného webu (viz kapitola 3.1.1). Jedná se o následující pohledy na data:

- přehled všech zpětných odkazů webu,
- grafy historického nárůstu/poklesu počtu zpětných odkazů a odkazujících domén,
- přehled nejčastěji použitých textů odkazů,
- přehled nejdůležitějších stránek webu podle příchozích odkazů,
- srovnání nově získaných vs. ztracených (odstraněných) odkazů,
- přehled unikátních odkazujících domén včetně počtu odkazů, které z těchto domén vedou.

Konkrétní využití funkce Site Explorer je popsáno v kapitole 3.1.8.

Všechny srovnávané nástroje zveřejňují základní čísla udávající velikost své databáze:

Majestic SEO nabízí dva různé indexy svých dat. Historický index obsahuje odkazy z méně často indexovaných (méně důležitých) stránek. Jeho aktualizace trvá přibližně měsíc. Čerstvý (Fresh) index obsahuje pouze odkazy ze stránek, které robot MajesticSEO prošel za poslední měsíc⁸.

Z přehledu není příliš zřejmý rozdíl mezi pojmy indexované stránky (pages crawled) a unikátní URL (unique URLs). Vysvětlení těchto údajů jsem bohužel nenalezl ani v nápovědě (16).

Tabulka 2: Informace o velikosti indexu služby Majestic SEO. (17)

Fresh index odkazy zaindexované v období 10. 1. 2012 až 10. 2. 2012 (poslední měsíc)	Historic index odkazy zaindexované v období 20. 6. 2006 až 21. 1. 2012 (starší jednoho měsíce)
• 18 609 493 195 indexovaných stránek • 109 142 624 816 unikátních URL	• 364 937 518 191 indexovaných stránek • 3 707 342 151 997 unikátních URL

Open Site Explorer využívá jako zdroj svých dat službu Linkscape firmy SEOMoz. I u něj existuje určité rozlišení mezi čerstvým indexem a indexem s historickými daty:

⁸ „At the time of writing, we are able to update our Historic Index monthly, and aim to update the Fresh Index daily“

Tabulka 3: Informace o velikosti indexu služby Open Site Explorer/Linkscope. (18)

„Čerstvá“ data odkazy zaindexované za posledních 82 dní	Celý index poslední aktualizace 17. 1. 2012
• 617 000 000 000 odkazů • 58 000 000 000 unikátních URL • 135 000 000 unikátních domén	• 6 000 000 000 000 odkazů • 580 000 000 000 unikátních URL • 135 000 000 unikátních domén

Ahrefs nerozlišuje indexy a nabízí tedy jen následující údaje o své databázi (19):

- **83 395 579 390** odkazů
- **11 743 056 308** unikátních indexovaných URL

Jak je vidět ze srovnání, každý z porovnávaných nástrojů uvádí trochu jiné informace o velikosti své databáze. Na základě těchto dat lze tedy nástroje porovnat jen velmi obecně. Dle udávané velikosti databáze vítězí MajesticSEO, následovaný službou Open Site Explore a Ahrefs.

3.1.3. Vlastní porovnání velikosti indexů odkazových databází

Abych ověřil oficiálně udávané statistiky velikosti databáze, otestuji jednotlivé databáze také vlastním jednoduchým testem. V něm srovnám jak **absolutní počet nalezených zpětných odkazů**, tak i **unikátních odkazujících domén**. Absolutní počet zpětných odkazů většinou může vypovídat o „hloubce“ indexu jednotlivých nástrojů, tj. jestli jsou schopni indexovat všechny stránky i u rozsáhlých webů. Počet unikátních domén zase ukazuje „šířku“ indexu, tj. jak velký počet různých webů je nástroj schopen indexovat.

U služby MajesticSEO jsem se rozhodl použít Fresh index místo History indexu. Právě tento index se čerstvostí dat blíží službě OpenSite Explorer. Mimo to mám zkušenost, že History index obsahuje velké množství odkazů, které již ve skutečnosti neexistují a History index tak pro mě nemá příliš velkou vypovídající hodnotu.

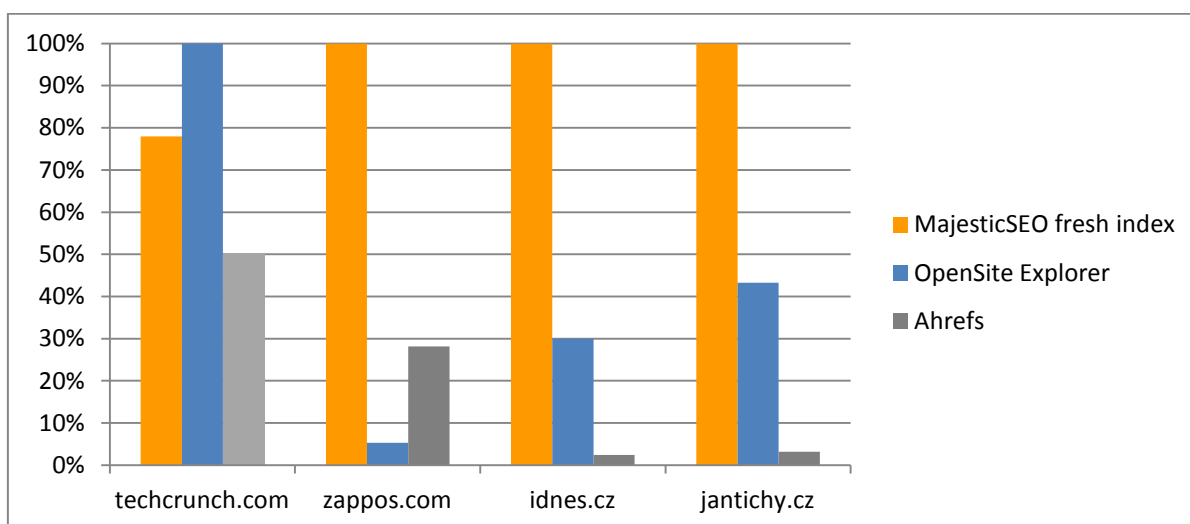
Testované weby:

- <http://www.techcrunch.com/> (magazín o nových technologických firmách)
- <http://www.zappos.com/> (americký e-shop s obuví)
- <http://www.idnes.cz> (české online periodikum)
- <http://www.jantichy.cz/> (menší blog v češtině)

Srovnání nalezených zpětných odkazů

Tabulka 4: Absolutní počty nalezených zpětných odkazů porovnávaných webů. (zdroj: autor)

	MajesticSEO fresh index	OpenSite Explorer	Ahrefs
<i>techcrunch.com</i>	2 358 479	3 024 254	1 525 110
<i>zappos.com</i>	7 796 360	414 867	2 196 041
<i>idnes.cz</i>	246 296	73 878	6 016
<i>jantichy.cz</i>	1 162	503	37

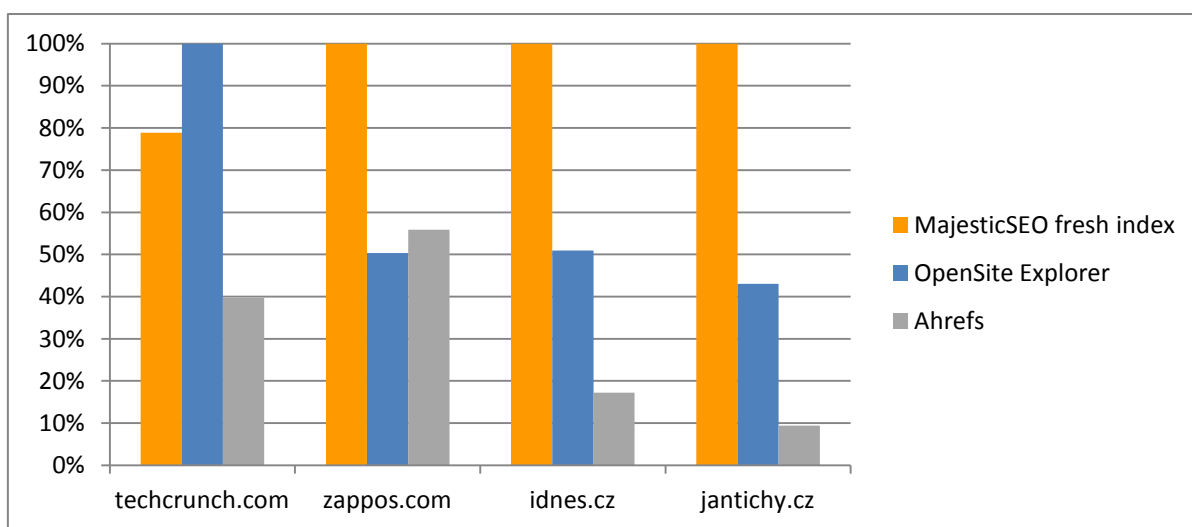


Obrázek 4: Relativní počet zpětných odkazů jak je odhalily zvolené odkazové databáze. Nástroj, který našel nejvíce odkazů, na grafu vždy představuje 100%. (zdroj: autor)

Srovnání nalezených odkazujících domén

Tabulka 5: Absolutní počty nalezených odkazujících domén porovnávaných webů. (zdroj: autor)

	MajesticSEO fresh index	OpenSite Explorer	Ahrefs
<i>techcrunch.com</i>	73 567	93 306	37 015
<i>zappos.com</i>	33 223	16 729	18 571
<i>idnes.cz</i>	4 861	2 475	837
<i>jantichy.cz</i>	202	87	19



Obrázek 5: Relativní počet nalezených odkazujících domén jak je odhalily zvolené odkazové databáze.

Nástroj, který našel nejvíce odkazů, na grafu vždy představuje 100%. (zdroj: autor)

Ze získaných dat vychází jako jasný vítěz služba MajesticSEO. Nalezla nejvíce odkazů i odkazujících domén u tří z celkových čtyř analyzovaných webů. Dobře použitelný může být i Open Site Explorer. Databáze služby Ahrefs je objektivně nejmenší, což je pravděpodobně způsobeno tím, že se jedná o velmi mladou službu.

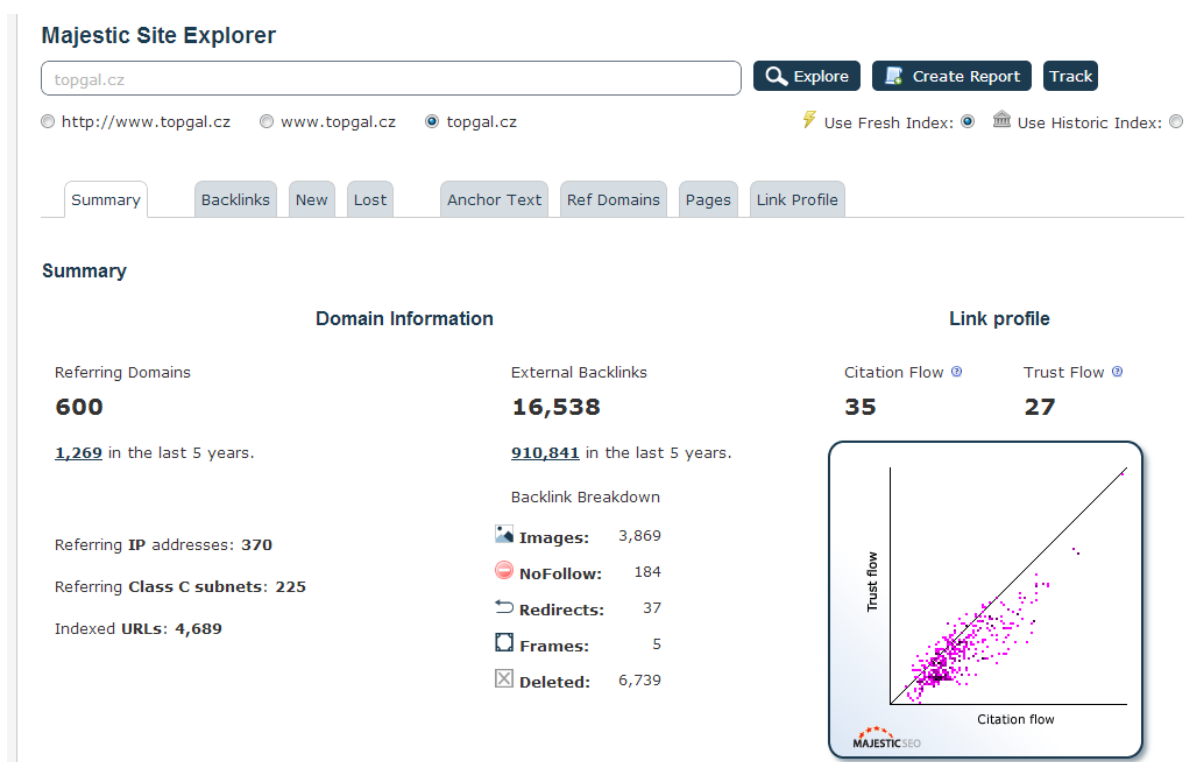
Grafy zobrazující **relativní rozdíly** mezi nástroji jsou téměř shodné pro absolutní počet odkazů i pro počet odkazujících domén. Pro komplexnější závěry by bylo vhodné rozšířit vzorek na např. 100 testovaných webů.

3.1.4. Kvalitativní srovnání nástrojů

MajesticSEO byla jedna z prvních odkazových databází na trhu. Nejen z toho důvodu je produkt nejvyspělejší po technologické stránce (největší a nejčastěji aktualizovaný index odkazů). Ze stejného důvodu jsou bohužel některé jejich nástroje vizuálně velmi zastaralé a složité na používání/pochopení.

Nástroje od MajesticSEO používám již několik let. Firma od počátku svého fungování bojuje o to, jak svá velice dobrá data uživatelům co nejlépe „prodat“. Rozhraní služby se sice průběžně vylepšuje, ale i dnes, téměř 4 roky po uvedení služby na trh je stále velmi složité, plné dlouhých tabulek napěchovaných daty.

Zjednodušeně se dá říci, že rozhraní MajesticSEO pouze poskytuje data, která služba neumí vizualizovat tak, aby odpovídala na problémy a dotazy uživatelů.



Obrázek 6: Ukázka rozhraní nástroje MajesticSEO Site Explorer. (zdroj: autor)



Obrázek 7: Ukázka rozhraní nástroje Ahrefs Site Explorer. (zdroj: autor)

Open Site Explorer od firmy SEOMoz patří spolu s MajesticSEO k nejstarším odkazovým databázím na trhu. Nástroj má velmi dobrý index odkazů (viz srovnání v předchozí kapitole 3.1.3), i když v poslední době naráží na problémy s jeho aktualizací (20).

Rozhraní Open Site Exploreru je také poměrně zastaralé, nabízí opět mnoho tabulek plných různých dat, ale pouze jediný graf (rozložení odkazujících domén podle koncovky – TLD).

Zásadní výhodou proti MajesticSEO je možnost pokročilého filtrování v zobrazovaných tabulkách dat a menší omezení zobrazovaných dat pro uživatele bezplatné verze nástroje.

Open Site Explorer je jediným z porovnávaných nástrojů, který nabízí statistiky také pro tzv. social metriky (počet sdílení stránky na Facebooku, Facebook „likes“, tvíty na Twitteru o stránce, zmínky na sociální síti Google+). Většina z těchto metrik bohužel v českém prostředí není příliš použitelná, takže jim nebudu věnovat více prostoru.

Unikátní a velice praktický je také nástroj porovnání statistik více webů mezi sebou⁹.

Poslední analyzovaná služba **Ahrefs** funguje teprve od června roku 2011¹⁰. Nabízí bezkonkurenčně nejlepší vizualizaci zobrazovaných dat. Mimo to přidává některé zajímavé funkce (např. poměry site-wide odkazů ke všem odkazům nebo výborná vizualizace odstraněných odkazů v čase). Rozhraní služby umožňuje velice pokročilé a hlavně jednoduché filtrování zobrazovaných dat (i když ne tak pokročilé jako u Open Site Exploreru).

Všemi těmito skvělými funkcemi se snaží vynahradit zatím nejmenší index zpětných odkazů mezi srovnávaným nástroji (viz kapitola 3.1.3).

Menší nevýhodou služby Ahrefs je (pochopitelný) fakt, že výrazně omezuje přístup pro neplatící uživatele. Registrovaný neplatící uživatel může každý den zobrazit maximálně 15 přehledů pro vybrané weby (přehled odkazů, odkazujících domén, textů odkazů atd.). Ostatní porovnávané nástroje žádné takovéto explicitní omezení nemají.

Ahrefs nabízí kromě Site Exploreru ještě další zajímavé nástroje (Site issues, Keyword Analysis), které zvyšují hodnotu této služby. Jejich popis je ale mimo rámec této práce.

3.1.5. Cenové srovnání vybraných nástrojů

Všechny zmíněné nástroje lze využívat za měsíční paušál nebo s omezeními zcela zdarma. I u bezplatného využívání je **vhodné se alespoň zaregistrovat**. Po registraci všechny nástroje navýší limity zobrazovaných zpětných odkazů, domén atd.

Zkušební období (trial) pro placené tarify nabízí pouze Open Site Explorer (30 dní po registraci zdarma; nutné zadat číslo kreditní karty). Ostatní nástroje tuto možnost

⁹ <http://www.opensiteexplorer.org/comparisons?site=www.topgal.cz%2F>

¹⁰ Viz archiv novinek na webu <https://ahrefs.com/news/index.php?p=3>

nenabízejí vůbec nebo jen v rámci speciálních propagačních akcí (např. MajesticSEO, viz (21)).

Níže uvedené ceny za placené tarify byly platné k **září roku 2012**.

Pokročilé funkce **MajesticSEO** lze začít využívat za **€39.99** (asi 1 000 Kč) měsíčně¹¹. Za tuto cenu dostane uživatel možnost používat MajesticSEO Site Explorer vytvářet pokročilé reporty včetně možnosti exportu čistých dat o odkazech pro maximálně 60 webů. Pro vlastní weby lze v MajesticSEO vytvářet pokročilé reporty zdarma.

Pro využívání MajesticSEO API je nutné zaplatit minimálně **€299.99** (asi 7500Kč měsíčně). I za tuto částku je ale možné API využívat jen pro interní použití.

SEOMoz Open Site Explorer je jedním z mnoha nástrojů firmy SEOMoz. Ke všem těmto nástrojům lze získat přístup za **\$99** (asi 2 000 Kč) měsíčně¹². Bezplatná verze tohoto nástroje nabízí dle mých zkušeností nejvíce dat a funkcí ze všech srovnávaných nástrojů.

Přístup k SEOMoz API je u tohoto tarifu velmi omezen (maximálně 1 požadavek každých 5 sekund) a pro masivnější použití se nehodí. Za využití API je nutné připlatit minimálně dalších \$500 (asi 10 000 Kč) měsíčně.

Tarify **Ahrefs** začínají na **\$79** měsíčně (asi 1 600 Kč) měsíčně¹³. V této ceně je zahrnuto až 2 000 denních požadavků na nástroj Ahrefs Site Explorer a 100 pokročilých reportů. Stejně jako u MajesticSEO lze pro vlastní weby vytvářet pokročilé reporty zdarma.

Také Ahrefs nabízí přístup ke svým datům pomocí API. To je lze využívat již v rámci nejnižšího tarifu zmíněného výše. V rámci tohoto tarifu lze například každý měsíc přes API získat počty odkazů pro 40 000 URL. Každý další dotaz na API bude stát 0.1 kreditu při ceně \$1 za 100 kreditů¹⁴.

¹¹ <https://www.majesticseo.com/plans-pricing>

¹² <http://www.seomoz.org/api/pricing>

¹³ http://ahrefs.com/pricing_plans.php

¹⁴ <https://ahrefs.com/api/>

U Ahrefs API není explicitně zakázáno využití pro třetí stranu jako u MajesticSEO. Firma ale v žádném případě nepovoluje zcela volné využití svých dat, viz podmínky užití API¹⁵.

3.1.6. Reálné kalkulace využívání API

Níže jsem se pokusil navrhnout několik případů užití API (pro zjednodušení pouze interní) a odhadované jednorázové poplatky. Případy užití vycházejí ze SEO analýz, které jsme prováděli ve firmě H1.cz.

Případ užití 1: Získáváme počty odkazů a odkazujících domén pro vybrané URL.

Vstupní data: 100 000 URL

Tabulka 6: Srovnání cen za využívání API pro Případ užití 1. (zdroj: autor)

MajesticSEO API	€299.99 (7 500 Kč)	Tarif <i>Platinum</i> Spotřebujeme 100 000 resources z 500 000, které tarif nabízí. Tarif Platinum je nejlevnější s přístupem k API.
SEOMoz API	\$500 (10 000 Kč)	Tarif <i>Low Volume</i> Umožňuje až 200 požadavků za sekundu, data budou tedy teoreticky získána za 500 sekund a můžeme získávat další.
Ahrefs API	\$1100 (22 000 Kč)	Využijeme 110 000 kreditů (Pro každou URL je nutné volat akci <code>get_backlinks_count</code> za 0.1 kreditů a <code>get_ref_domains_ips_count</code> 1 kredit).

Případ užití 2: Získáváme stejná data jako v prvním případě a k tomu ještě počty odkazů/odkazujících domén obsahující v textu odkazu konkrétní obecné klíčové slovo (např. letenky). Vstupní data: 100 000 URL

¹⁵ „If you want to show our data to the third-party, don't start integration without conditioning the rules of usage with us.“ viz <https://ahrefs.com/api/>

Tabulka 7: Srovnání cen za využívání API pro Příklad užití 2. (zdroj: autor)

MajesticSEO API	€1199.96 (30 000 Kč)	Tarif <i>Majestic 1000</i> Spotřebujeme 100 000 000 resources ¹⁶ z 200 000 000, které tarif nabízí.
SEOMoz API	\$500 (10 000 Kč)	Tarif <i>Low Volume</i> Potřebujeme 3 požadavky pro každou URL, tedy 3 000 požadavků. Data budou tedy teoreticky získána za 15 sekund.
Ahrefs API	nelze získat počet odkazujících domén	

3.1.7. Kterou odkazovou databází zvolit?

Na základě porovnání z předchozích kapitol jsem připravil tabulku, srovnávající přednosti a nedostatky analyzovaných nástrojů:

MajesticSEO

- + Největší a nejčastěji aktualizovaná databáze odkazů
- + Nejlevnější placený tarif z porovnávaných služeb
- Složitě na používání
- Data z API je možné využívat jen pro interní použití
- Komplikované účtování využívání API

¹⁶ 100 000 URL * 1000 AnalysisResUnits per URL = 100 000 000 AnalysisResUnits (zdroj: <http://developer-support.majesticseo.com/api/commands/>)

SEOMoz včetně Open Site Explorer

- | | |
|--|--|
| + Velmi dobré možnosti filtrování dat | – Menší velikost indexu |
| + Nejlepší data u bezplatné verze | – Problémy s aktualizací indexu odkazů |
| + Přehledné a jasné tarify služby | – Horší vizualizace odkazových dat |
| + Nástroj na porovnávání více webů | |
| + Nejkomplexnější API vhodné i pro složitější operace | |
| + Nízká cena za využívání API při větším počtu požadavků | |

Ahrefs

- | | |
|---|---|
| + Jednoduché tarify i rozhraní | – Velmi omezená bezplatná verze |
| + Výborné filtrování a vizualizace dat | – Nejasné možnosti využití dat z API pro třetí strany |
| + Obsahuje další zajímavé SEO nástroje v ceně | – Až příliš jednoduché API bez pokročilých funkcí |
| + Nová služba, která se velmi rychle rozvíjí | |

Pro **začínající a středně pokročilé uživatele** bych doporučil nástroj SEOMoz Open Site Explorer od SEOMozu. Pokud ale Ahrefs zlepší velikost svého indexu odkazů, dal bych přednost tomuto nástroji.

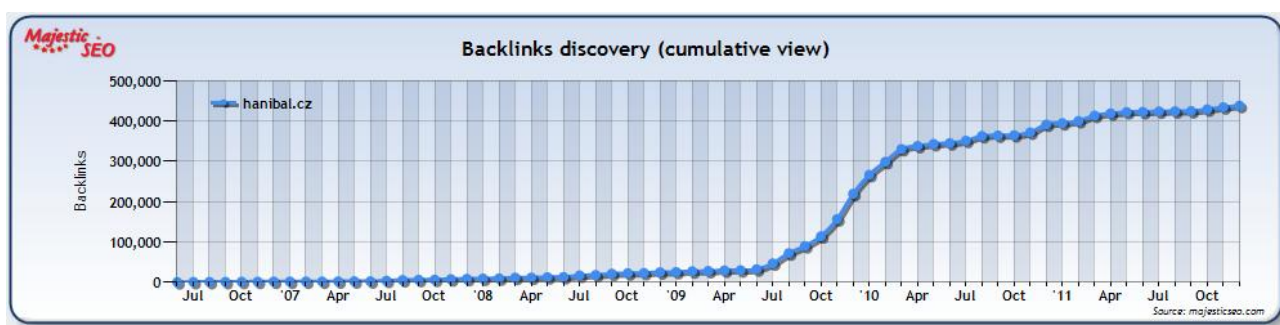
Pokročilým uživatelům doporučuji využívání SEOMoz Open Site Exploreru díky pokročilým možnostem filtrování. Pokud je nezajímají předpřipravené přehledy dat a chtějí si vytvořit vlastní, sáhnou pravděpodobně k exportu čistých dat z MajesticSEO.

Firma **poskytující SEO konzultace a zároveň budující vlastní interní SEO nástroje** využije nejlépe API SEOMozu, které v ideálním případě může kombinovat s daty z MajesticSEO.

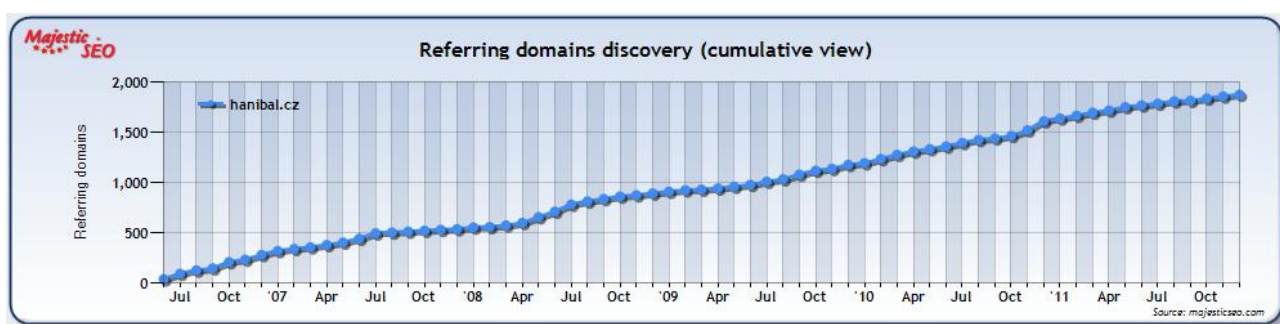
Firmy, které staví vlastní veřejné SEO nástroje, by měly vyzkoušet SEOMoz API, které je dle mého názoru funkčně i cenově nejzajímavější. U ostatních nástrojů je využití dat pro veřejné použití problematické.

3.1.8. Praktické využití: analýza konkurence

S pomocí databází zpětných odkazů lze jednoduše zjistit, z jakých webů konkurence získává odkazy, případně jaké typy odkazů získávají. K porovnávání využijeme nástroj Backlink History a web www.hanibal.cz.



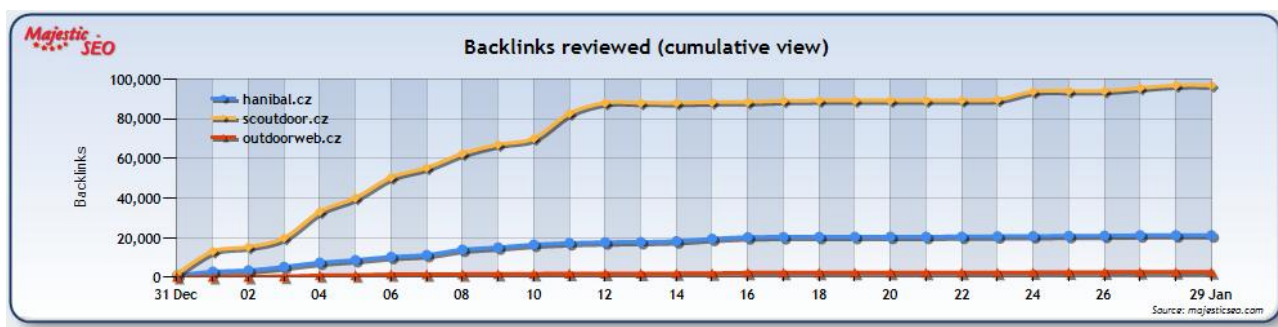
Obrázek 8: Vývoj celkového počtu odkazů webu hanibal.cz v čase pomocí MajesticSEO. (zdroj: autor)



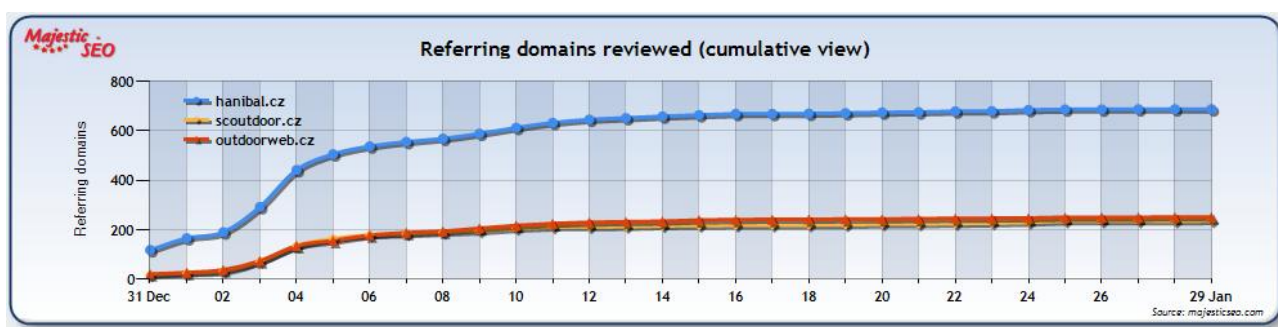
Obrázek 9: Vývoj počtu unikátních domén (viz začátek kapitoly 3.1), které odkazují na web hanibal.cz v čase s pomocí MajesticSEO. (zdroj: autor)

Z uvedených grafů je zřejmé, že absolutní počet odkazů webu Hanibal.cz rostl v období mezi červencem 2010 a dubnem 2011 mnohem rychleji, než počet unikátních odkazujících domén. Vysoký počet odkazů na jednu odkazující doménu ukazuje, že správce webu se v daném období pravděpodobně uchýlil k nákupu tzv. site-wide odkazů. To jsou odkazy, které se zobrazují na všech stránkách webu, nejčastěji v patičce.

Nástroj Backlink History je ještě lepší použít pro srovnání více webu mezi sebou. Níže uvedené grafy nám pomůžou vyhodnotit, jak dobře nebo špatně na to náš web je relativně v porovnání s konkurencí:



Obrázek 10: Vývoj celkového počtu odkazů webu hanibal.cz v čase pomocí MajesticSEO. (zdroj: autor)



Obrázek 11: Vývoj počtu unikátních domén, které odkazují na web hanibal.cz v čase s pomocí MajesticSEO. (zdroj: autor)

S využitím databází zpětných odkazů lze také odhadnout, na jaké fráze se daný web ve vyhledávacích soustředí. Stačí např. s pomocí Open Site Exploreru seskupit odkazy podle jejich textů a následně texty seřadit podle počtu odkazů, ve kterých jsou obsaženy.

Z výsledků Open Site Exploreru je zřejmé, že web hanibal.cz se soustředí především na následující fráze (22):

- **termoprádlo** (4 804 odkazů, 29 odkazujících domén),
- **crocs** (1 138 odkazů, 19 odkazujících domén),
- **spacíky** (8 449 odkazů, 17 odkazujících domén),
- **sněžnice** (2 294 odkazů, 15 odkazujících domén).

Velké množství odkazů obsahuje také název samotného webu (hanibal), což značí, že Hanibal je silnou a známou značkou.

Kontrolu konkurence stačí provádět např. čtvrtletně. Je vhodnější svůj čas věnovat raději zlepšování vlastních produktů, služeb a způsobů jejich propagace.

3.2. Nástroje pro analýzu klíčových slov

Uživatelé ve vyhledávačích hledají stejné věci různými slovy i různými tvary slov.

Analýza klíčových slov je proces, kdy hledáme co největší množství různých klíčových slov, která:

- **jsou tematicky příbuzná s obsahem našeho webu**, charakterizují sortiment našeho e-shopu atd.,
- **uživatelé je používají při hledání ve vyhledávačích** (čím častěji, tím lépe),
- **jsou co nejméně konkurenční** (tj. existuje co nejmenší počet webů, které mají ve svých stránkách daná slova obsažena; hodně konkurenční slova mohou na náš web přivést velké množství návštěvníků, ale je velmi složité na tato slova získat dobrou pozici ve vyhledávačích),
- **pomáhají plnit cíle našeho webu** (např. přivádějí na náš web návštěvníky z vyhledávačů, kteří u nás následně nakoupí, přečtou si náš článek, atd.),
- **jsou obecná i konkrétní** (texty stránek našeho webu by měly být maximálně variabilní a obsahovat obecná i konkrétní klíčová slova, např. *letenky* vs. *levné letenky do new yorku v lednu 2012*).

Analýza klíčových slov by měla v ideálním případě předcházet každému vytváření textů pro web. Pomáhá nám především:

- ujasnit si správnou strukturu našeho webu,
- psát texty pro náš web tak, aby následně získaly maximální návštěvnost z vyhledávačů,
- zjistit, o které zboží/služby z našeho sortimentu se lidé zajímají a o které nikoliv.

Konkrétní postupy budou popsány v navazujících podkapitolách.

3.2.1. Popis vybraných nástrojů pro analýzu klíčových slov

Analyzované nástroje: Google AdWords Keyword Tool, Nástroj pro návrh klíčových slov Skliku, Wordtracker, SEMRush

Nástroje pro návrh klíčových slov nám pomáhají při analýze klíčových slov nezapomenout na žádnou důležitou skupinu slov ale i konkrétní slovní spojení. Většinou fungují tak, že do nich vložíme jedno základní klíčové slovo, podle kterého nám poradí další slova příbuzná, synonyma atd., která lidé používají při hledání ve vyhledávačích.

Google AdWords KeywordTool je celosvětově nejznámějším a nejpoužívanějším nástrojem pro návrh klíčových slov. To je pravděpodobně způsobeno tím, že jej Google nabízí zcela zdarma. Google jej vytvořil především pro účely svého reklamního systému AdWords, ale dá se velmi dobře používat i pro SEO.

Ač se jedná o nástroj mezinárodní, je velmi dobře použitelný i pro češtinu.

Nejjednodušeji jej lze vyzkoušet na adrese

<https://adwords.google.com/select/KeywordToolExternal>. Bez přihlášení do účtu

Google AdWords ale nástroj zobrazuje **pouze omezené výsledky** (max. 50 klíčových slov). Pro jeho plnou funkčnost je vhodné si zdarma zaregistrovat účet Google AdWords na adrese <https://adwords.google.com> a nástroj používat právě v rámci tohoto účtu.

Nástroj umožňuje pro jedno hledání vložit více výchozích slov. Dle mých zkušeností ale pro dosažení nejlepších výsledků je vhodné do nástroje vkládat vždy pouze jedno klíčové slovo.

Google AdWords Keyword Tool nezobrazí úplně všechna příbuzná klíčová slova v rámci jednoho hledání. Je vhodné s ním experimentovat a postupně do něj zadávat stále konkrétnější slova (*stany -> rodinné stany -> rodinné stany pro 5 osob*).

Pro Česko je vhodné používat nástroj s následujícím nastavením:

- oblast a jazyky: **Česká republika, všechny jazyky** (s nastavením Česká republika, čeština nástroj u většiny slov nevrátí žádnou hledanost, jedná se pravděpodobně o chybu na straně Google),
- Zobrazit návrhy a statistiky pro: **Stolní počítače a notebooky**,
- Typ shody: **přesná** (chceme zjistit údaje pouze pro zadané klíčové slovo, ne pro slova příbuzná a synonyma)

Nástroj pro návrh klíčových slov Skliku je součástí reklamního systému Sklik českého vyhledávače Seznam.cz. Nabízí méně funkcí, než nástroj od Googlu, ale zato vrací

k zadanému klíčovému slovu často mnohem více relevantních klíčových slov. V nástroji Skliku nelze oproti nástroji Googlu klíčová slova seřadit podle hledanosti.

Pro používání nástroje je nutná bezplatná registrace do Skliku na stránce <http://www.sklik.cz>. Při vyhledávání klíčových slov nástroj nikde ve svém rozhraní nezobrazuje celkový počet nalezených klíčových slov. Pro zjištění této informace je potřeba po kliknutí na tlačítko „Hledej klíčová slova“ analyzovat HTTP komunikaci mezi naším prohlížečem a serverem Skliku. Celkový počet nalezených slov odhalí atribut *count* XML elementu *keywords*.

Wordtracker nabízí několik nástrojů pro internetové marketéry pracující se SEO i PPC reklamou. Jejich nástroj pro návrh klíčových slov **WordTracker Keywords** je možné po registraci na 7 dní vyzkoušet zcela zdarma.

Za produktem **SEMRush** stojí SEOquake Team, který vytvořil stejnojmenné populární rozšíření pro SEO do prohlížeče Firefox. Nástroj především umožňuje zobrazit odhadovanou návštěvnost a umístění ve vyhledávacích téměř libovolného většího webu. Data pro své analýzy získává analýzou výsledků hledání Google a Bingů na asi 80 000 000 klíčových slov.

SEMRush nabízí databáze klíčových slov podle konkrétních zemí (např. US, UK, DE, FR a další). Česko v seznamu bohužel chybí. V neplacené verzi zobrazuje pouze prvních 10 nalezených klíčových slov, takže relevanci výsledků je velmi těžké určit.

3.2.2. Vlastní porovnání

Jako porovnávací metriku jsem zvolil celkový **počet klíčových slov**, která porovnávané nástroje vyhledají po zadání vybraného obecného klíčového slova.

Testovací klíčová slova:

- letenky
- slevy
- batohy
- bílé zboží

- iphone

Tabulka 8: Počty nalezených klíčových slov s využitím popsaných nástrojů. (zdroj: autor)

Výchozí klíčové slovo	Nástroj Google AdWords	Nástroj Skliku	WordTracker	SEMRush (Related keywords report)
letenky	127	3 157	83	79
slevy	67	14 038	2	0
batohy	105	8 268	7	0
nokia	800	24 072	3 548	5 998
iphone	800	6 057	8 454	19 752

Z hlediska hlavní porovnávané metriky **počet vyhledaných klíčových slov** jednoznačně zvítězil nástroj Skliku, který vrací řádově více klíčových slov. Je tak vhodný pro opravdu rozsáhlé analýzy klíčových slov. Vzhledem k chybějící možnosti řadit klíčová slova podle hledanosti je vhodné ještě využívat nástroj Googlu. U obou nástrojů je relevance vrácených klíčových slov velmi dobrá. Nástroj Skliku vrací oproti nástroji Google mnohem větší počet long-tailových slov (např. *letenka praha tenerife*).

WordTracker a **SEMRush** jsem do přehledu zařadil spíše pro ilustraci. Pro česká klíčová slova nenavrhnu téměř žádná příbuzná klíčová slova a nejsou proto vhodné pro reálné využití.

3.2.3. Praktické využití: návrh struktury webu

Struktura e-shopů je i v dnešní době stále většinou navrhována programátory nebo v nejlepším případě majiteli daného e-shopu, kteří často nemají velké zkušenosti se SEO. Proto navrhuji strukturu webu pouze podle vlastních pocitů a názorů. Nejvhodnějším řešením je přitom opět využití analýzy klíčových slov, která pomůže navrhnout:

- **nejvhodnější pojmenování sekcí zboží,**

Používají potenciální zákazníci e-shopu spíše termín „hokejky“ nebo „hokejové hole“? Znají termíny, které jsou použity v popiscích zboží? Nechápu významy některých výrazů odlišně? Analýza tyto otázky pomůže zodpovědět.

- **nejpřirozenější kategorizaci zboží.**

Neobsahuje web zbytečné kategorie, které zákazníci nepoužívají? Nechybí naopak kategorie/filtrování, které by zákazníci ocenili (např. filtrování hokejek podle délky, materiálu čepele atd.). Nejsou důležité kategorie zanořeny příliš hluboko?

Tato kapitola obsahuje metodiku analýzy klíčových slov pro účel návrhu vhodné struktury webu.

Při analýze je vhodné začít výběrem tématu webu. U menších webů může tématem být zaměření celého webu, pro větší weby je vhodné si analýzu rozdělit např. po kategoriích webu nebo jiných ohraničených tematických celcích. V následujícím příkladu bude analyzován větší e-shop se sportovním vybavením. Analýza je zaměřena výhradně na sekci s potřebami pro lední hokejisty.

Nejvhodnějším postupem před samotnou analýzou pomocí nástrojů pro návrh klíčových slov je vytvoření základního seznamu důležitých klíčových slov. V tomto seznamu se nacházejí slova, která člověka napadnou při pohledu na sortiment a strukturu analyzované sekce webu. Pro hokejové vybavení jsou to např. slova:

- hokejky,
- brusle,
- puky,
- hokejové chrániče,
- helmy,
- suspenzory,
- hokejové chrániče kolen,
- potřeby na hokej,

- vybavení na hokej.

Pro prvotní analýzu je vhodné použít nástroj pro návrh klíčových slov Google AdWords. Na základě zadaného klíčového slova *hokejky* nástroj Googlu našel 51 příbuzných klíčových slov v jednotném i množném čísle. Všechna nalezená slova je poté nutné rozdělit do skupin dle jejich tématu, typu, obecnosti vs. konkrétnosti atd.

Pro slova vycházející ze základního slova *hokejky* mohou vzniknout např. následující skupiny:

- **značky** (hokejky bauer, hokejky warrior),
- **hledání s určením místa** (hokejky praha),
- **varianty zboží** (kompozitová hokejka, brankářská hokejka),
- **konkrétní modely** (hokejka easton s19),
- **překlepy a chybně/nestandardně napsaná slova** (hokejky warior, floorball hokejky),
- **příslušenství a příbuzné vybavení** (čepele na hokejky, brusle, grip na hokejku),
- **zajímavá synonyma** (hokejové hole),
- **slova mimo téma našeho webu** (hokejky na florbal, inline brusle bauer).

Podobný postup se aplikuje i na další klíčová slova z úvodního výběru. Relevantní nalezená slova je nutné si průběžně poznamenávat např. do Excelu spolu s následujícími charakteristikami:

- **hledanost**,
 - ▶ Sloupec **Celosvětový objem vyhledávání za měsíc** udává přibližný dvanáctiměsíční průměr počtu uživatelských dotazů pro dané klíčové slovo ve vyhledávání Google.
 - ▶ Sloupec **Místní objem vyhledávání za měsíc** udává přibližný dvanáctiměsíční průměr počtu uživatelských dotazů pro dané klíčové slovo ve vyhledávání Google pouze v České republice.

- ▶ U ryze českých klíčových slov se stačí orientovat podle sloupce **Celosvětový objem vyhledávání za měsíc**. Pokud některá z analyzovaných klíčových slov jsou mezinárodní (nike, iphone atd.), je vhodné sledovat při práci s klíčovými slovy také veličinu **Místní objem vyhledávání za měsíc**.

- **konkurenčnost.**

Kolik konkurenčních webů se na dané klíčové slovo zobrazuje ve výsledcích hledání vyhledávače.

Výsledný seznam slov (**161** klíčových slov) je k dispozici v souboru **hokejove-potreby-kw.xlsx**, umístěném na CD přiloženém k této práci. Pro účel návrhu struktury je potřeba, aby slova byla seřazena sestupně podle hledanosti od nejobecnějších (nejhledanějších) slov k těm nejkonkrétnějším (nejméně hledaným). Na základě tohoto seřazení vznikne např. následující návrh struktury hokejové sekce:

1. úroveň: hlavní stránka kategorie zboží

Klíčová slova: *hokejové vybavení, hokejové potřeby, hokejová výstroj*

Skupinu je vhodné doplnit ještě o úroveň kombinující název hlavní stránky kategorie zboží a značku výrobce, např.: *hokejové vybavení bauer, hokejová výstroj warrior*.

2. úroveň: hlavní podkategorie zboží

Klíčová slova: *hokejky brusle, hokejové chrániče, hokejové rukavice, hokejové kalhoty*

Opět je vhodné doplnit ještě úroveň kombinující název podkategorie zboží a značku výrobce, např.: *hokejky warrior, brusle bauer*. U podkategorie se doplnění o značku většinou řeší přidáním filtru zboží, který zobrazí pouze zboží vybrané značky (viz ilustrační obrázky pod tímto odstavcem).

Brusle

Díky naší široké nabídce hokejových bruslí získáte vždy **nejlepší poměr ceny a výkonu!** Nakupte u nás brusle pro všechny věkové i výkonnostní kategorie hokejistů.



Obrázek 12: Sekce „Brusle“ e-shopu Sportobchod.cz. Žádný filtr značky není vybrán. (zdroj: autor)

Brusle Bauer

Díky naší široké nabídce hokejových bruslí získáte vždy **nejlepší poměr ceny a výkonu!** Nakupte u nás brusle pro všechny věkové i výkonnostní kategorie hokejistů.



Obrázek 13: Sekce „Brusle“ e-shopu Sportobchod.cz. V sekci je zobrazen pouze podvýběr bruslí značky Bauer. (zdroj: autor)

Z obrázku je patrné, že pro klíčové slovo „brusle bauer“ vznikla nová stránka webu zaměřená pouze na brusle této značky. Vyhledávač bude uživatele hledající klíčové slovo „brusle bauer“ a slova příbuzná posílat především na tuto konkrétnější stránku, nikoliv na obecnou hlavní stránku sekce „Brusle“. Je to způsobeno především tím, že stránka má klíčové slovo „brusle bauer“ použito v titulku a hlavním nadpisu stránky.

3. úroveň: upřesněné podkategorie zboží

Klíčová slova: *kompozitové hokejky, brankářské hokejky*

I v tomto případě je vhodné doplnit úroveň kombinující název upřesněné podkategorie zboží a značku výrobce, např.: *kompozitové hokejky warrior, brankářské hokejky bauer*

Finální návrh struktury sekce hokejových potřeb je k dispozici v souboru **struktura_hokejove_sekce.png** na CD přiloženém k této práci.

3.2.4. Praktické využití: Porovnání značek

Nástroj pro návrh klíčových slov Google lze velmi dobře použít např. pro porovnávání hledanosti různých značek. V následujícím příkladě jsem zkusil porovnat hledanost leteckých společností ČSA, Wizzair, Ryanair a Easyjet. Do nástroje jsem zadal všechny kombinace, které by uživatelé mohli hledat včetně překlepů:

- čsa
- wizzair
- wiz air
- vizair
- viz air
- wizzair
- wizz air
- ryanair
- ryan air
- easyjet
- easy jet

Tabulka 9: Výsledná data o hledanosti vybraných klíčových slov. (zdroj: autor)

Klíčové slovo	Celosvětový objem vyhledávání za měsíc	Místní objem vyhledávání za měsíc
čsa	5 400	4 400
české aerolinie	1 300	1 000
české aerolinky	73	85
czech airlines	135 000	4 400
wizair	40 500	480
wiz air	9 900	170
vizair	2 900	36
viz air	2 400	73
wizzair	1 830 000	14 800
wizz air	450 000	5 400
ryanair	13 600 000	18 100
ryan air	1 220 000	1 600
easyjet	7 480 000	12 100
easy jet	1 500 000	3 600

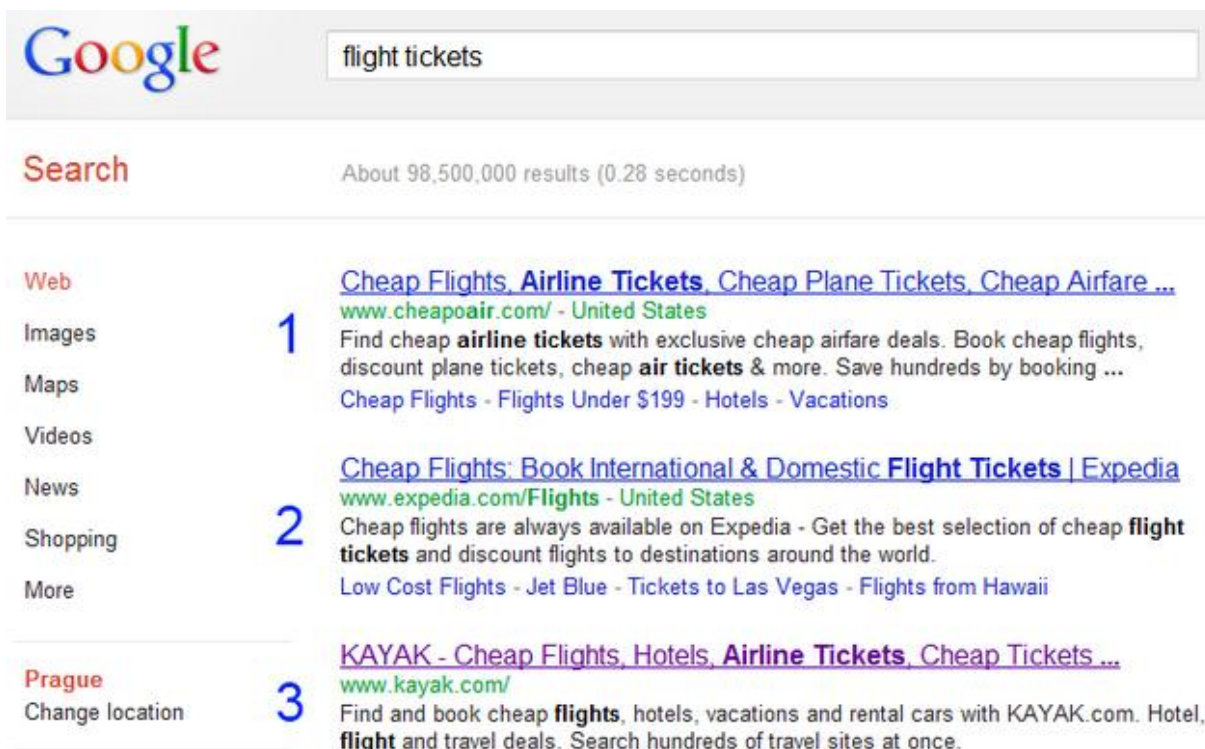
Tabulka 10: Hledanost je vhodné pro lepší přehlednost seskupit podle značky letecké společnosti. (zdroj: autor)

Klíčové slovo	Celosvětový objem vyhledávání za měsíc	Místní objem vyhledávání za měsíc
čsa české aerolinie české aerolinky czech airlines	141 773	9 885
wizair wiz air vizair viz air wizzair wizz air	2 335 700	20 959
ryanair ryan air	14 820 000	19 700
easyjet easy jet	8 980 000	15 700

Z tabulky vyplývá, že z vybraných leteckých společností je celosvětově nejhledanější RyanAir, následovaný EasyJet a WizzAir. ČSA skončily v žebříčku poslední. Při omezení pouze na Česko je hledanost WizzAir a RyanAir téměř shodná, následuje EasyJet a na závěr opět ČSA.

3.3. Nástroje pro měření pozic ve vyhledávačích

Při měření pozic zjišťujeme, na které pozici ve výsledcích vybraného vyhledávače na vybrané klíčové slovo se náš web nachází.



Obrázek 14: Výsledky vyhledávání na klíčové slovo *flight tickets*. Hledaný web *kayak.com* má ve výsledcích na toto klíčové slovo třetí pozici na *google.com*. (zdroj: autor)

Měřit pozice svého webu ve vyhledávačích dnes lze pomocí stovek různých nástrojů. Nástroje jsou většinou velice podobné, odlišují se většinou pouze v následujících parametrech:

- vyhledávače, na kterých se pozice měří (Google.cz, Google.com, Seznam.cz, Yahoo.com)
- periodičita měření (každou hodinu, denně, jednou za týden atd.),
- maximální počet sledovaných výsledků (např. do 60 výsledků).

Vysoký počet nástrojů je dán tím, že získávání dat o pozicích je technicky mnohem méně náročné, než např. vytvoření databáze zpětných odkazů z celého Internetu (viz databáze zpětných odkazů v kapitole 3.1.).

3.3.1. Jak se pozice ve vyhledávačích měří?

Vyhledávače nenabízejí své výsledky v žádné strojově zpracovatelné formě (XML, JSON atd.) zdarma a bohužel ani za peníze. Nemají zájem na tom, aby jejich výsledky někdo vytěžoval a snažil se tak např. odhalit algoritmus, jakým jsou výsledky řazeny.

Všechny nástroje měřící pozice, včetně těch od renomovaných firem jako SEOMoz, musí proto výsledky hledání získávat parsováním standardních výsledků ve formátu HTML (tzv. scraping).

3.3.2. Jak se scraping provádí?

V našem příkladu budeme provádět scraping stránky výsledků (**SERP** – Search Engine Result Page) vyhledávače Seznam.cz. Stránka výsledků vyhledávání Seznamu má URL

```
http://search.seznam.cz/?q={keyword}&count=20&from={from}&blindFriendly=1
```

Co znamenají jednotlivé parametry?

- **q** – hledané klíčové slovo,
- **count** – počet záznamů na stránku,
Seznam standardně zobrazuje 10 výsledků na stránku. Za předpokladu, že potřebujeme zkontrolovat pozice v prvních 60 výsledcích, museli bychom stáhnout 6 stránek výsledků. Pokud ale nastavíme stránkování po 20 záznamech, stačí nám stáhnout pouze 3 stránky výsledků Seznamu (počet požadavků klesne 2x).
- **from** – parametr stránkování; from=21 znamená, že Seznam má přeskočit prvních 20 výsledků a zobrazit výsledky od 21. pozice dále.
- **blindFriendly=1** – vynutit zobrazení stránky výsledků pro zrakově postižené,
Verze výsledků pro zrakově postižené je datově mnohem menší, a proto pro scraping vhodnější.

Získaný HTML kód stránky s výsledky se zpracovává např. pomocí regulárních výrazů.

```

<h3> <a href="http://www.letenky.com/" class="resultLink" title="Přejít na:
http://www.letenky.com/" data-dot="title"> 0 &mdash; <b>Letenky</b>.com - levné <b>
letenky</b> online do celého světa </a> </h3>

<h3> <a href="http://www.letenky.cz/" class="resultLink" title="Přejít na:
http://www.letenky.cz/" data-dot="title"> 1 &mdash; <b>Letenky</b>, <b>Letenky</b>.cz,
levné <b>letenky</b> online </a> </h3>

<h3> <a href="http://www.kralovna.cz/" class="resultLink" title="Přejít na:
http://www.kralovna.cz/" data-dot="title"> 2 &mdash; <b>Letenky</b>, Královna, <b>
letenky</b> online </a> </h3>

```

Obrázek 15: HTML kód prvních tří výsledků Seznam.cz, který parsujeme pomocí regulárních výrazů. (zdroj: autor)

Pro získání URL adres všech výsledků na stránce použijeme následující jednoduchý regulární výraz:

```
<a href="([^\"]+)" class="resultLink" title="Přejít na:[^\"]+">
```

Pokud systém používá scraping, je samozřejmě nutné co nejčastěji (alespoň jednou denně) kontrolovat, že regulární výraz je stále aktuální a scraping výsledků probíhá správně. Popis tohoto testování je mimo téma této práce.

3.3.3. Problémy se stránkováním

Některé vyhledávače jako např. Google řadí své výsledky odlišně při použití standardního stránkování po 10 výsledcích a než při stránkování po 60 výsledcích.

Toto chování lze ověřit např. vyhledáním slova **antiradary** na Google.cz:

AntiRadary.NET - největší prodejce s antiradary pro ČR !!
www.antiradary.net/
Antiradary Beltronics & AntiLaser & Escort Radar, vše SKLADEM, **antiradary** za nejlepší ceny! Info: 731 800 400, 731 800 200.
[Stránka Google+](#)

 Na Sekyrce 1392/2 160 00 Praha
 731 800 500

[Antiradary diskuse - Kontakt](#)

Antiradary - Antiradary, radarové detektory, laserové rušičky
www.antiradary.eu/
Antiradary - antiradar, radarové detektory, laserové rušičky, Beltronics i Valentine One.

Pasivní antiradar a antiradary prodej montáž a recenze. Laserové ...
www.cz-antiradary.cz/
 Prodej a montáž **antiradarů**, **antiradar**, Valentine One V1, detektor radarů, gps, voice, beltronics, rx65, 975e. Recenze **antiradaru** V1 CZ.

Obrázek 16: První tři výsledky při stránkování po 10 výsledcích. (zdroj: autor)

AntiRadary.NET - největší prodejce s antiradary pro ČR !!
www.antiradary.net/
Antiradary Beltronics & AntiLaser & Escort Radar, vše SKLADEM, **antiradary** za nejlepší ceny! Info: 731 800 400, 731 800 200.
[Stránka Google+](#)

 Na Sekyrce 1392/2 160 00 Praha
 731 800 500

[Antiradary diskuse - Kontakt](#)

Antiradar Beltronics RX65e EURO - CZ | AntiRadary.NET
www.antiradary.net/antiradar-rx65-euro-beltronics/
 Vítěz testu - Beltronics RX65e Euro CZ. Antiradar Beltronics RX65e EURO CZ je modelem roku 2007 a vyznačuje se prakticky nulovou poruchovostí, vysokou ...

Antiradary a legálnost - legálnost antiradarů | AntiRadary.NET
www.antiradary.net/antiradary-legalnost-zakon/
 Antiradary a legálnost - legálnost antiradarů. 5.5.2011: Sněmovna schválila zásadní snížení pokut za použití aktivního antiradaru. Dlouho očekávaná novela ...

Obrázek 17: První tři výsledky při stránkování po 60 výsledcích. (zdroj: autor)

Ze screenshotů je vidět, že Google v případě zobrazení 60 výsledků na jednu stránku k sobě seskupuje výsledky ze stejných domén. Důvod tohoto chování se mi bohužel nepodařilo zjistit.

3.3.4. Co je potřeba brát v úvahu při získávání výsledků vyhledávačů

Některé vyhledávače (např. Google) zobrazují přihlášeným uživatelům trochu jiné (tzv. personalizované) výsledky. Personalizace výsledků v některých případech závisí také na geografické poloze uživatelského počítače.

Při měření pozic je proto nutné personalizaci vypnout přidáním parametru **pws=0** do Google SERP URL. Vhodné je také explicitně stanovit zemi, ze které uživatel vyhledává pomocí parametru **gl** (např. **gl=CZ** pro Česko).

Geografická poloha našeho počítače také určuje, které datové centrum Googlu nám odpoví. Pokud na svém počítači zavolám v příkazové řádce příkaz *ping www.google.cz*, cílová IP adresa je 173.194.35.184. Pokud stejný příkaz zavolám z našeho serveru umístěném v Irsku, cílová IP adresa je 74.125.24.94.

Rozdílné IP adresy představují různá datová centra Google. Data mezi datovými centry se neustále synchronizují. Přesto se čas od času stane, že různá datová centra vrací mírně odlišné výsledky.

3.3.5. Právní aspekty scrapingu

Vyhledávač Google zakazuje scraping (ač velmi vágně) přímo ve svých pravidlech užití: *„Naše služby nesmíte zneužívat. Např. nesmíte do služeb zasahovat ani se do nich pokoušet získat přístup jiným způsobem než pomocí určeného rozhraní a podle dodaných pokynů.“*(23).

Pokud Google usoudí, že se nějaká osoba nebo software snaží o scraping jeho výsledků, odepře (většinou dočasně) další přístup k vyhledávání z dané IP adresy.

Our systems have detected unusual traffic from your computer network. Please try your request again later. [Why did this happen?](#)

This page appears when Google automatically detects requests coming from your computer network which appear to be in violation of the [Terms of Service](#). The block will expire shortly after those requests stop.

This traffic may have been sent by malicious software, a browser plug-in, or a script that sends automated requests. If you share your network connection, ask your administrator for help — a different computer using the same IP address may be responsible. [Learn more](#)

Sometimes you may see this page if you are using advanced terms that robots are known to use, or sending requests very quickly.

IP address: 188.95.60.135

Time: 2012-08-19T07:09:13Z

URL: http://173.194.70.106/search?

hl=cs&num=10&source=hp&q=rotoped&btnG=Vyhledat+Googlem&lr=&aq=f&oq=&pws=0&ie=UTF-8&oe=UTF-8&as_qdr=all&gl=CZ

Obrázek 18: Hlášení, které Google zobrazí při odepření přístupu k vyhledávání. (zdroj: autor)

Podobné blokování automatizovaných dotazů jako Google používají i všechny ostatní vyhledávače včetně Seznam.cz, Yahoo, Bing a dalších.

Odepření přístupu ze strany Google většinou lze zabránit využitím následujících postupů (24):

- posíláním dotazů z velkého množství (čím víc tím lépe) různých IP adres,
- dostatečným čekáním mezi jednotlivými požadavky (např. náhodné čekání mezi 15 a 60 sekundami),
- zasíláním různých user-agentů (simulace různých uživatelů s různými prohlížeči).

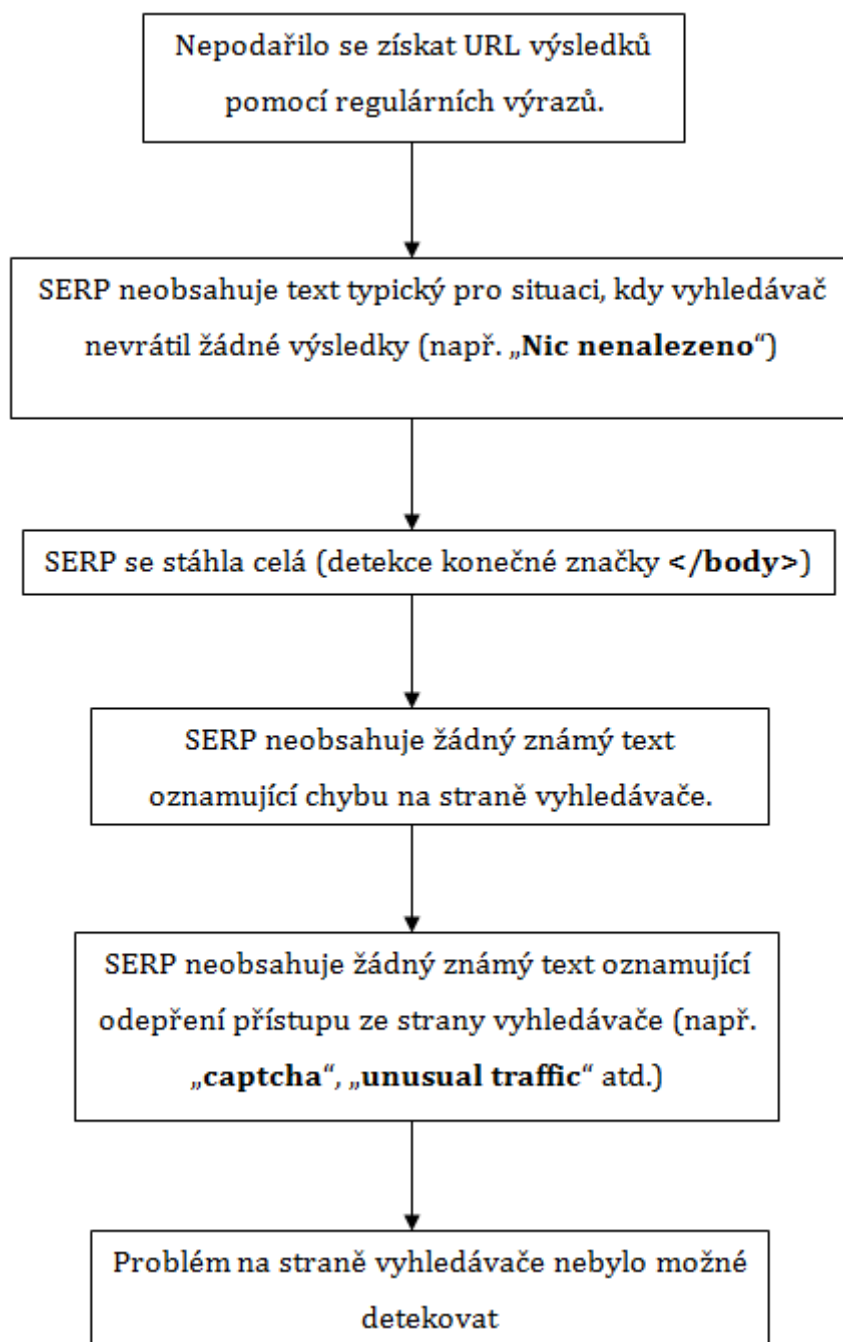
Na závěr je potřeba říci, že Google sice ve svých podmínkách zakazuje scraping svých výsledků a snaží se mu zabránit i přímo odepřením přístupu k výsledkům z „příliš aktivní“ IP adresy. Na druhou stranu Google s mnoha firmami, kterého jeho výsledky takto „zneužívají“ spolupracuje a toto chování u nich toleruje.

Google by samozřejmě mohl žalovat všechny stovky, možná tisíce firem, které jeho podmínky porušují. Výsledek takových žalob by mu ale nic zásadního nepřinesl a zcela jistě by mu pokazil jméno v očích všech SEO konzultantů.

3.3.6. Detekce problémů a nečekaných situací na stráně vyhledávače

Vyhledávač má, stejně jako každý jiný softwarový produkt, své chyby. Pokud bychom pro získávání výsledků komunikovali s vyhledávačem pomocí nějakého standardizovaného API, dostali bychom v případě problémů pravděpodobně konkrétní, strojově zpracovatelné chybové hlášení. To v případě scrapingu většinou neplatí a chyby na straně vyhledávače je nutné detekovat jiným způsobem.

Situace, které je nutné řešit:



Obrázek 19: Schéma detekce problémů vyhledávače při měření pozic. (zdroj: autor)

4. Korelační analýza výsledků vyhledávače

Seznam.cz

V praktické části své diplomové práce se zaměřím na korelační analýzu výsledků vyhledávání vyhledávače Seznam.cz. Metodika měření vychází z podobné analýzy *2011 Search Engine Ranking Factors*, zveřejněné na serveru SEOMoz.org pro vyhledávače Google (25). Finální analýzu jsem provedl o mezi 3. 8. 2012 – 5. 8. 2012.

4.1. Cíle

Spočítat korelace mezi jednotlivými **SEO faktory, které ovlivňují pozice v SERP na Seznam.cz** a **skutečnými pozicemi** ve výsledcích hledání Seznamu. Z těchto vztahů následně rozhodnout, které SEO faktory jsou pro dosažení dobrých pozic pravděpodobně důležitější a které méně důležité.

Dalším cílem je ověření nebo vyvrácení hypotéz definovaných v kapitole 4.4. Ty vycházejí jak z důležitostí jednotlivých faktorů podle průzkumu na SEOFaktory.cz, tak z mých vlastních názorů získaných při dvouleté praxi SEO konzultanta.

4.2. Výběr atributů (SEO faktorů) korelace

Faktory jsem vybíral podle průzkumu zveřejněném na webu SEOFaktory.cz (26). Autoři tohoto webu se zeptali vybraných českých SEO konzultantů na nejdůležitější faktory, které ovlivňují umístění webových stránek ve výsledcích vyhledávání.

V průzkumu každý z porotců dostal k ohodnocení všechny faktory. U každého faktoru označil jeho důležitost v rozmezí 0 (žádný vliv) až 100 (maximální vliv). Pokud porotce faktor nechtěl ohodnotit, tak jeho hodnocení mohl přeskočit, aby zbytečně nezkresloval data. Výsledné hodnocení každého faktoru představuje aritmetický průměr hodnocení jednotlivých porotců.

Do své analýzy jsem zahrnul **pouze on-page SEO faktory** z průzkumu, které:

- dosáhly v průzkumu SEOFaktory.cz důležitosti **alespoň 30 %**,
- byly **strojově zpracovatelné** a strojová analýza daného faktoru byla realizovatelná v rozumném čase (např. pro analýzu faktoru *Klíčové slovo použito v*

anchor textu interního odkazu na stránce bych musel vytvořit robota, který by proindexoval celý daný web a zaznamenal interní odkazy; vytvoření takového robota by zabralo měsíce práce; pro faktor *Stáří (doba) od vytvoření stránky* jsem zase nenašel vhodný zdroj dat),

- měly **jasnou definici**.

(např. faktor *Existence rozsáhlého, unikátního obsahu stránky* je vágně popsán a proto jsem pro něj nemohl provést korelační analýzu).

Off-page SEO faktory jsem do korelační analýzy nakonec mohl zahrnout jen velmi omezeně, protože jsem pro výpočet jejich korelace neměl vhodná data. Především jsem postrádal podrobná data o zpětných odkazech webů.

Tato data nabízí firmy MajesticSEO, SEOMoz a Ahrefs, zmíněné kapitole 3.1. V době provádění korelační analýzy (červenec a srpen 2012) nabízela všechna potřebná data pouze služba MozScape API od firmy SEOMoz.

Přístup k tomuto API je bohužel velice drahý, měsíční paušál začínal na 500 dolarech, viz kapitola 3.1.6. Pokusil jsem se SEOMoz emailem požádat o omezený přístup k API pouze pro účely této práce, nedostalo se mi ale bohužel odpovědi.

Z dřívější doby jsem měl přístup alespoň k omezenému **MajesticSEO Light API**. Toto API je zdarma, ale poskytuje pouze informace o absolutním počtu odkazů vybraného webu, počtu odkazujících domén a stránek. Právě metriky *absolutní počet odkazů* a *počet odkazujících domén* jsem nakonec do své analýzy zařadil, viz kapitola 4.3 práce.

4.3. Vybrané atributy korelace

V této kapitole najdete seznam SEO faktorů, které jsem nakonec použil pro svou korelační analýzu.

Hlavní faktory jsou vybrány na základě průzkumu ze SEOFaktory.cz, dle postupu zmíněného v předchozí kapitole.

Doplňkové faktory doplňují faktory hlavní. Např. faktor *Klíčové slovo použito v názvu domény* (např. www.klicoveslovo.cz) bylo potřeba vydefinovat jasněji a proto jsem jej rozdělil na hlavní faktor *Klíčové slovo se shoduje s doménou druhého řádu (přesná shoda)* a doplňkové faktory (viz kapitola 4.3.4):

- klíčové slovo se shoduje s doménou druhého řádu (volná shoda),
- klíčové slovo je obsaženo v hostname,
- počet výskytů klíčového slova v hostname.

Mezi doplňkové faktory jsem zařadil i faktory *Počet odkazů webu* a *Počet unikátních webů, které odkazují na web*, protože na SEOFaktory.cz nejsou nikde takto přesně zmíněny.

4.3.1. Titulek stránky

U slov v titulku stránky byly odstraněny předložky a spojky. Při porovnávání záleželo na pořadí slov.

Klíčové slovo použito kdekoliv v tagu <title>	
<i>Upřesněný název:</i>	Klíčové slovo použito kdekoliv v tagu <title> (přesná shoda)
<i>Popis:</i>	Příklad: levne letenky levne = 1 výskyt fráze "levne letenky" Z toho faktoru jsem odvodil doplňkový faktor Klíčové slovo použito kdekoliv v tagu <title> (volná shoda) , ve kterém nezáleželo na pořadí slov zdrojového klíčového slova v titulku stránky.
<i>Důležitost:</i>	SEOFaktory.cz: 79 %

Klíčové slovo použito jako první slovo v tagu <title>	
<i>Důležitost:</i>	SEOFaktory.cz: 71 %

4.3.2. Nadpisy stránky (H1 – Hx)

Umístění klíčového slova jsem vyhodnocoval vždy na úrovni jednoho nadpisu. Nikoliv tedy spojením všech textů nadpisů dohromady a následnému vyhodnocení.

Klíčové slovo použito kdekoli v tagu nadpisu <h1>	
<i>Upřesněný název:</i>	Klíčové slovo je obsaženo v prvním tagu H1
<i>Popis:</i>	Z toho faktoru jsem odvodil doplňkový faktor Počet výskytů klíčového slova v prvním tagu H1.
<i>Důležitost:</i>	SEOFaktory.cz: 53 %

Klíčové slovo použito jako první slovo/a v tagu nadpisu <h1>	
<i>Upřesněný název:</i>	Klíčové slovo je první v prvním tagu H1
<i>Důležitost:</i>	SEOFaktory.cz: 51 %

Klíčové slovo použito kdekoli v tagu dalších nadpisů <h2> - <h6>	
<i>Popis:</i>	Počet výskytů napříč nadpisy <h2> - <h6>. Odstraněny předložky, spojky, nezáleží na pořadí slov.
<i>Důležitost:</i>	SEOFaktory.cz: 44 %

4.3.3. Atributy ostatního obsahu stránky

Použití klíčových slov / množství opakování klíčových slov v HTML stránky	
<i>Upřesněný název:</i>	Počet výskytů klíčového slova na stránce
<i>Popis:</i>	Počet výskytů v obsahu stránky - přesná shoda, bez HTML tagů a jejich atributů; kontrolována také verze slova s odstraněnými předložkami.
<i>Důležitost:</i>	SEOFaktory.cz: 39 %

Klíčové slovo použito v prvních 50 - 100 slovech v HTML kódu stránky	
<i>Upřesněný název:</i>	Počet výskytů klíčového slova v prvních 100 slovech stránky
<i>Popis:</i>	Počet výskytů v prvních 100 slovech stránky očištěných od HTML tagů; kontrolována také verze slova s odstraněnými předložkami.
<i>Důležitost:</i>	SEOFaktory.cz: 34 %

4.3.4. Atributy domény a URL

Klíčové slovo použito v názvu domény (např. www.klicoveslovo.cz)	
<i>Upřesněný název:</i>	Klíčové slovo se shoduje s doménou druhého řádu (přesná shoda)
<i>Popis:</i>	Přesné pořadí slov, odstraněny předložky, spojky, diakritika; pouze pro domény 2. řádu Z toho faktoru jsem odvodil doplňkové faktory: • Klíčové slovo se shoduje s doménou druhého řádu (volná shoda), • Klíčové slovo je obsaženo v hostname, • Počet výskytů klíčového slova v hostname, (levne letenky levne = 1 výskyt "levne letenky")
<i>Důležitost:</i>	SEOFaktory.cz: 50 %

Kód země v koncovce domény (např. .cz, .co.uk, .de, .fr, .sk, atd.)	
<i>Upřesněný název:</i>	1. Stránka má českou (.cz) doménu 2. Stránka má jinou než českou doménu
<i>Důležitost:</i>	SEOFaktory.cz: 34 %

4.4. Hypotézy

1) Nejvyššího korelačního koeficientu dosáhne některý ze SEO faktorů, který popisuje umístění klíčového slova v tagu <title> stránky

Popis	SEO faktory popisující umístění klíčového slova v titulku stránky dosáhly v hodnocení na SEOFaktory.cz nejvyšší důležitosti. Proto předpokládám, že míra korelace bude u nich nejvyšší. Faktory: • Klíčové slovo použito kdekoli v tagu <title> (důležitost 79 %) • Klíčové slovo použito jako první slovo v tagu <title> (důležitost 71 %)
-------	--

2) Faktor *Počet výskytů klíčového slova na stránce* se umístí mezi 3 faktory s nejnižším korelačním koeficientem

Popis	Mnoho SEO konzultantů si myslí, že text optimalizované stránky by měl mít konkrétní hustotu klíčového slova, na které je stránka optimalizována (např. 3 – 5 %). Já si naopak myslím, že na hustotě vůbec nezáleží a je dostatečné, když optimalizovaná stránka obsahuje důležité klíčové slovo jen jednou nebo dvakrát. Předpokládám tedy, že korelační koeficient tohoto faktoru se bude blížit 0 nebo bude dokonce záporný.
-------	--

3) Faktory popisující umístění klíčového slova v tagu <title> budou mít vyšší korelační koeficient než faktory popisující umístění klíčového slova v hlavním nadpisu stránky <h1>

Popis	V hodnocení SEOFaktory.cz dosáhly faktory popisující umístění klíčového slova v tagu <title> vyšší důležitosti (79 % a 71 %), než faktory popisující umístění klíčového slova v hlavním nadpisu stránky <h1> (53 % a 51 %). Proto předpokládám, že budou také lépe korelovat s pozicí stránky v SERP.
-------	---

4) Některý z faktorů, který popisuje umístění klíčového slova v doméně nebo hostname, bude mezi třemi faktory s nejvyšším korelačním koeficientem

Popis	Přítomnost klíčového slova v doméně (např. www.letenky.cz) způsobuje, že zpětné odkazy vedoucí na tuto doménu často obsahují ve svém textu klíčové slovo <i>letenky</i>. Lidé totiž většinou odkazují prostým zkopírováním URL webu z adresního řádku prohlížeče. Předpokládám, že tento jev způsobí vysoký korelační koeficient u faktorů z této skupiny.
-------	--

4.5. Výběr klíčových slov pro analýzu

Vlastním firmu, která vytváří a spravuje český webový SEO nástroj Collabim, který mimo jiné umožňuje měření pozic klíčových slov ve vyhledávačích. Jako vstupní data korelační analýzy jsem proto použil pozice a klíčová slova klientů právě z Collabimu.

Vybraná klíčová slova se týkají mnoha různých oborů lidské činnosti. Pro výběr slov jsem použil následující podmínky:

- analyzujeme jen pozice na vyhledávači Seznam.cz,
(*WHERE searchEngine = 2*),
- vybíráme obecnější klíčová slova, která mají hledanost v Česku (svGoogleLocal) větší než 250 hledání měsíčně,
(*WHERE collabim.keywords.country = "CZ" AND svGoogleLocal > 250*); číslo 250 jsem vybral tak, aby analyzovaná skupina obsahovala kolem 10 000 klíčových slov (stejně jako v případě analýzy SEOMozu),
- vybíráme pouze aktuální výsledky naměřené za poslední 2 dny,
(*WHERE serp_results.added >= (CURDATE() - INTERVAL 1 DAY)*)
- kvůli zjednodušení nás zajímají pouze fráze spojené z maximálně 3 slov,
(počítáme počet mezer ve frázi – *WHERE substrCount(text, " ") <= 2*)
- opět kvůli zjednodušení vybíráme pouze fráze bez nealfanumerických znaků, (čárky, apostrofy, uvozovky atd.; - *WHERE keywords.text REGEXP '^[:alnum:]+[[:punct:]]+\$*)

Podoba použitého SQL dotazu pro výběr klíčových slov:

```
INSERT IGNORE INTO
    correlation_keywords (keywordId, text, svGoogle,
svGoogleLocal)
SELECT
    keywordId,
    keywords.text,
    keywords.svGoogle,
    keywords.svGoogleLocal
FROM
    serp_results
    JOIN collabim.keywords ON (serp_results.keywordId =
keywords.id)
WHERE
    searchEngine = 2
    AND collabim.keywords.country = "CZ"
    AND svGoogleLocal > 250
    AND serp_results.added >= (CURDATE() - INTERVAL 1 DAY)
    AND substrCount(text, " ") <= 2
    AND keywords.text REGEXP '^[:alnum:] ]+$';
```

Uvedený SQL dotaz vložil do tabulky *correlation_keywords* v MySQL databázi celkem **11 363 klíčových slov**.

Struktura celé databáze je zobrazena v **příloze 1 této práce**.

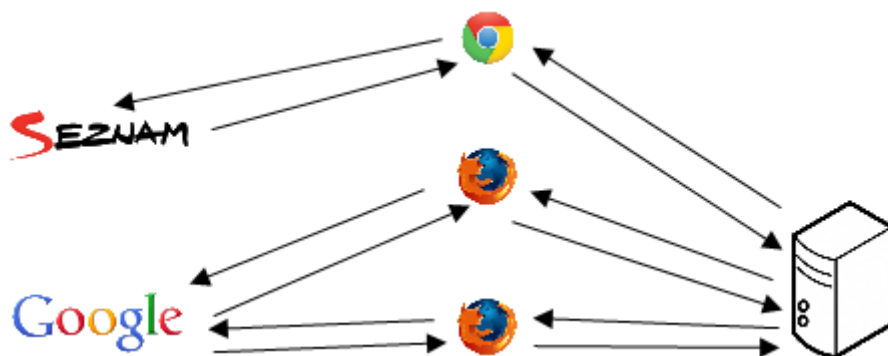
Tabulka 11: Distribuce hledanosti vybraných klíčových slov. (zdroj: autor)

Měsíční lokální hledanost slova na Google	Počet klíčových slov
<250; 400>	4 266
<401; 600>	2 174
<601; 1000>	2 209
<1001; 5000>	2 173
<5001; 10 000>	734
10 001+	200

4.6. Jak Collabim získává data o pozicích webů ve vyhledávačích?

Obecný způsob získávání těchto dat je popsán v kapitole 3.3. Collabim využívá pro získávání výsledků Googlu a Seznamu rozšíření, která si jeho uživatelé dobrovolně instalují do svých prohlížečů (Google Chrome, Firefox a Safari). Dotazy na vyhledávače jsou tak posílány z mnoha různých IP adres a prohlížečů s různým nastavením.

Tímto postupem je zajištěno, že nám vyhledávače neodepřou přístup ke svým výsledkům. Díky využívání našich uživatelů jako prostředníků nemají šanci dotazy z Collabimu detekovat.



Obrázek 20: Schéma získávání výsledků z vyhledávačů Google a Seznam. (zdroj: autor)

4.7. Zpracování výsledků

K vybraným klíčovým slovům jsem do tabulky *corralation_results* vložil celkem **227 260 výsledků** (URL) z vyhledávače Seznam.cz. Pro každé klíčové slovo jsem pracoval s prvními **20 výsledky hledání** z vyhledávače Seznam.cz. Tento vyhledávač neumožňuje žádnou personalizaci výsledků, takže výsledky hledání jsou vždy shodné pro všechny uživatele.¹⁷

Proces analýzy jsem rozdělil na **3 části**:

1. stažení obsahu stránek, které se objevily ve výsledcích hledání Seznam.cz,
2. výpočet hodnot všech SEO faktorů pro jednotlivé stránky,
3. výpočet korelačního koeficientu pro jednotlivá klíčová slova,
4. výpočet souhrnného korelačního koeficientu SEO faktorů.

Pro zpracování dat jsem využil instanci (virtuální server) nainstalovaný v cloudu Amazon Web Services. Aplikace pro zpracování dat byla napsána v jazyce PHP (Nette 2 + Zend Framework 1). Většinu funkcionality jsem pokryl jednotkovými testy napsanými ve frameworku PHPUnit 3.6.

Použité knihovny v jazyce PHP jsou k dispozici na přiloženém CD.

4.7.1. Stažení obsahu stránek, které se objevily ve výsledcích hledání Seznam.cz

Stahovací robot se hlásil pod standardním user agentem prohlížeče Firefox 4. Maximální čas čekání na stažení stránky (timeout) byl 30s. Pokud se robotovi nepodařilo stáhnout stránku napoprvé, zkusil to následně ještě dvakrát s pětivteřinovým čekáním.

Po stažení jsem stránku překódoval z původního kódování na UTF-8. Pokud byla velikost stránky větší než 1 MB, zpracovával jsem pouze první 1 MB dat stránky. Výsledek tohoto zpracování jsem uložil do databáze (tabulka *correlation_pages*).

¹⁷ „A co Seznam? U toho se bát nemusíme, i když s nějakou personalizací experimentuje, výsledky pro různé uživatele zůstávají zatím stejné.“ (27)

Zpracování (stažení) všech 227 260 stránek a jejich uložení zabralo serveru asi **20 hodin času**. Z uvedeného počtu se z různých důvodů nepodařilo stáhnout pouze 703 stránek. Velikost výsledné databáze byla **13,6 GB** (z toho 5,5 GB tvořily databázové indexy).

4.7.2. Výpočet hodnot všech SEO faktorů pro jednotlivé stránky

Analýzy vybraných 23 SEO faktorů pro všech 227 260 stránek zabralo serveru asi **40 hodin času**. V následujících bodech popíšu, jak jsem pracoval s konkrétními částmi zpracovávaných stránek.

HTML obsah stažených stránek jsem před analýzou zpracoval a normalizoval tak, aby ve výsledné analýze bylo co nejméně chyb. Chybou v tomto případě myslím jak chybu technického rázu (např. špatné zpracování českých znaků v regulárních výrazech), tak chybu ve smyslu odlišnosti mého algoritmu zpracování od reálného algoritmu vyhledávače Seznam.cz.

Použité normalizační operace:

- **převedení na malá písmena,**

Vyhledávač Seznam.cz nerozlišuje v hledaných frázích velikost písmen.¹⁸

- **odstranění diakritiky,**

Tato operace byla nutná např. při porovnávání klíčového slova (může obsahovat diakritiku) s hostname/doménou konkrétního webu (neobsahuje diakritiku).

- **odstranění předložek a spojek,**

Při výpočtu počtu výskytů jsem jako vstup použil jak původní klíčové slovo, tak klíčové slovo s odstraněnými předložkami a spojkami. Tímto chováním jsem zajistil, že např. klíčové slovo *letenky nairobi* bude detekováno v textu „Kupte si levné letenky do nairobi“.

Tato normalizace pravděpodobně není zcela shodná se skutečným normalizačním procesem vyhledávače Seznam.cz. Ten místo odebírání předložek a spojek zjišťuje

¹⁸ „Dotazy lze zadávat jak s diakritikou tak bez ní, nezáleží ani na velikosti písmen, proto dotaz „olympiada“ i „Olympiáda“ vrátí stejné výsledky.“(28)

blízkost (tzv. proximitu) slov v textu a porovnává ji s blízkostí slov v hledané frázi. Tento postup je ale algoritmicky náročný a nebylo v mých silách jej naimplementovat.

Konkrétní znaky a slova, které byly odebrány, lze nalézt v **příloze 2** práce.

- **odstranění HTML značek,**

Pomocí PHP funkce `strip_tags()` a převedení HTML entit na běžné znaky jsem ze vstupu odebral zbytečné informace. Vyhledávače sice dávají klíčovým slovům, uzavřeným např. v HTML značce `<strong>` vyšší význam. V rámci definovaného zadání ale tento problém nebylo potřeba řešit, takže jsem ho ve prospěch zjednodušení dalšího zpracování záměrně ignoroval. Text nadpisů (značky `<h1>` až `<hX>`) byl samozřejmě zpracován zvlášť.

- **normalizace bílých znaků,**

Především došlo k odstranění zalomení řádků, odebrání nadbytečných mezer mezi slovy, nahrazení tabulátorů a jiných nestandardních oddělovačů slov za mezery. Touto operací jsem v kombinaci s odstraněním HTML značek z velice složitého HTML kódu na vstupu získal jednoduše zpracovatelný čistý text. Bez této operace bylo vyhledávání frází o více slovech v textu stránky mnohem komplikovanější a náchylnější na vznik chyb.

- **odstranění hlavičky HTML stránky (obsah mezi `<head>` a `</head>`).**

Tuto operaci jsem provedl, abych oddělil uživateli viditelný text stránky a meta informace uvedené v HTML hlavičce stránky (ty jsou dále zpracovávány samostatně).

Uvedené normalizační operace jsem aplikoval při analýze SEO faktorů následujícím způsobem:

Titulek stránky:

1. výběr obsahu HTML značky `<title>`,
2. převedení na malá písmena,
3. odstranění HTML značek,
4. normalizace bílých znaků.

Nadpisy stránky (H1 – Hx):

1. odstranění hlavičky HTML stránky,
2. převedení na malá písmena,
3. odstranění HTML značek,
4. normalizace bílých znaků.

Atributy ostatního obsahu stránky:

1. odstranění hlavičky HTML stránky,
2. převedení na malá písmena,
3. odstranění HTML značek,
4. normalizace bílých znaků,
5. výběr prvních 100 slov z obsahu (pouze o SEO faktoru *Počet výskytů klíčového slova v prvních 100 slovech stránky*).

Odkazy:

1. převedení na malá písmena.

Atributy domény a URL

1. převedení na malá písmena,
2. odstranění diakritiky,
3. odstranění předložek a spojek.

4.7.3. Výpočet korelačního koeficientu pro jednotlivá klíčová slova

Po vzoru zmíněné analýzy na serveru SEOMoz.org jsem pro výpočet korelačního koeficientu použil Spearmanův koeficient pořadové korelace. Tomuto typu výpočtu korelace jsem dal přednost před Pearsonovým korelačním koeficientem především proto, že analyzované faktory jsou velice různorodé a jejich rozdělení není normální (Gausovo).

$$\rho = 1 - \frac{6 \sum_i (p_i - q_i)^2}{n(n^2 - 1)}$$

Obrázek 21: Vzorec pro výpočet Spearmanova koeficientu pořadové korelace. (29)

Při analýze jsem, po vzoru analýzy na serveru SEOMoz, vypočítal korelační koeficient pro každý SEO faktor každého klíčového slova zvlášť.

V jazyce PHP jsem nenalezl žádnou použitelnou knihovnu pro výpočet Spearmanova korelačního koeficientu. Tuto knihovnu jsem si musel tedy naimplementovat sám.

Nejprve bylo nutné vytvořit algoritmus, který seřadí vstupní parametry (pozice a hodnoty každého jednotlivého SEO faktoru) a přiřadí jim **pořadový rank**.

Pro ilustraci uvedu výstup řadícího algoritmu pro prvních 10 výsledků vyhledávání na klíčové slovo **domácí pekárny**. Vstupem pro řazení byly atributy:

- **pozice na klíčové slovo** (v tabulce jako „Pozice“),
- **počet výskytů klíčového slova na stránce** (v tabulce jako „Počet KW“).

Tabulka 12: Přehled ohodnocených výsledků vyhledávání na klíčové slovo „domácí pekárny“. (zdroj: autor)

URL	Pozice	Rank pro pozice	Počet KW	Rank pro počet KW
http://domacipekarny.dama.cz	1	10	1	9
http://www.pekarny.unas.cz	2	9	12	4
http://domaci-pekarny.heureka.cz	3	8	12	4
http://www.mall.cz/domaci-pekarny/	4	7	4	8
http://www.mojepekarna.cz/domaci-pekarny	5	6	8	6
http://www.nakupka.cz/bila-technika/domaci-pekarny/	6	5	15	2
http://www.domaci-pekarny.cz/pekarny/	7	4	0	10
http://eta.czdomaci-pekarny	8	3	7	7
http://www.mimibazar.cz/recepty.php?id=30	9	2	18	1
http://domaci-pekarny.elektromedia.cz	10	1	12	4

O výpočet pořadového ranku se postaral algoritmus popsáný v **Příloze 3**. Při vytváření algoritmu jsem se inspiroval stránkami (30) a (31).

Největší výzvou při implementaci bylo zvládnout správně operace s čísly s dlouhým desetinným rozvojem. Standardní datový typ float (plovoucí desetinná čárka) jazyka PHP nebylo možné použít, protože umožňuje práci pouze se 14 desetinnými místy¹⁹ s limitovanou přesností (32).

Pro přesnější práci s čísly nabízí PHP knihovny **BC Math** a **GMP** (GNU Multiple Precision) (33). GMP pracuje pouze s celými čísly, proto jsem nakonec využil knihovnu BC Math (přesnost na 1000 desetinných míst, pevná desetinná čárka).

¹⁹ The size of a float is platform-dependent, although a maximum of $\sim 1.8e308$ with a precision of roughly 14 decimal digits is a common value

Implementace knihovny pro výpočet Spearmanova korelačního koeficientu s využitím BC Math byla ve výsledku relativně jednoduchá:

```
class SpearmanCorrelationOfBigNumbersCounter {

    public function count(array $array1, array $array2) {
        bcscale(1000);

        $n = count($array1);

        $top = bcmul('6', $this->sumOfSquares($array1, $array2));

        $bottom = bcsub(bcpow($n, '3'), $n);

        return bcsub('1', bcddiv($top, $bottom));
    }

    private function sumOfSquares(array $array1, array $array2) {
        $sum = '0';

        for ($i = 0; $i < count($array1); $i++) {
            $p = $array1[$i];
            $q = $array2[$i];

            $difference = bcsub($p, $q);
            $poweredDifference = bcpow($difference, 2);

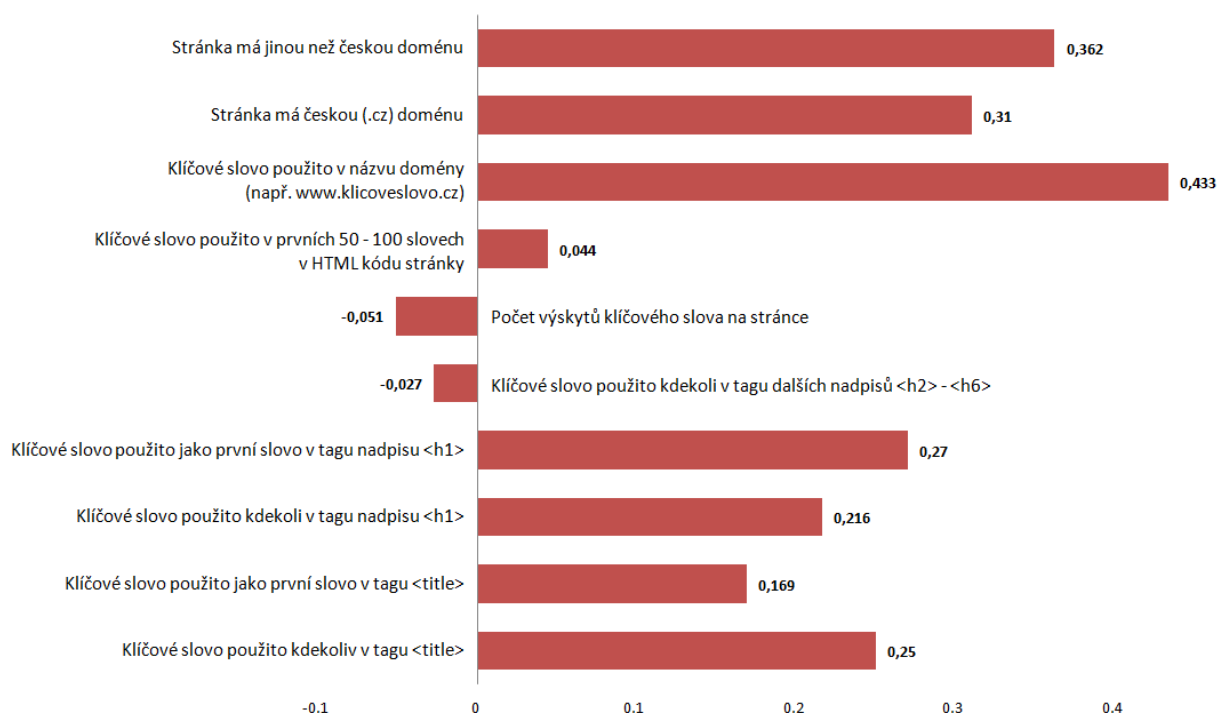
            $sum = bcadd($sum, $poweredDifference);
        }

        return $sum;
    }
}
```

Spearmanův korelační koeficient jsem spočítal **pro každé klíčové slovo zvlášť** (tedy vždy z 20 výsledků pro daný SEO faktor). Výpočet korelačních koeficientů pro všech 23 SEO faktorů u 227 260 výsledků vyhledávání zabralo serveru **3 hodiny času**.

4.7.4. Korelační koeficienty pro jednotlivé SEO faktory

V poslední fázi analýzy jsem dle metodiky SEOMoz spočítal průměrný korelační koeficient každého SEO faktoru napříč všemi klíčovými slovy. Výsledek ilustruje graf níže:



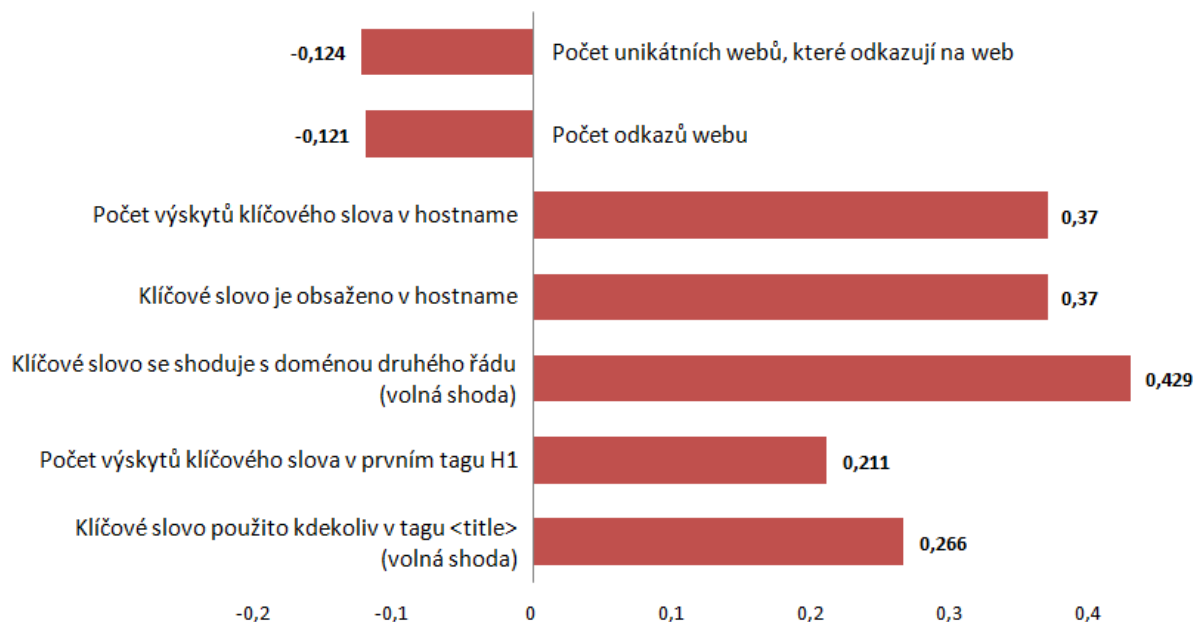
Obrázek 22: Graf vypočítaných korelačních koeficientů pro definované SEO faktory. (zdroj: autor)

Téměř shodná míra korelace faktorů *Stránka má jinou než českou doménu* a *Stránka má českou (.cz) doménu* napovídá, že na typu domény prvního řádu (TLD) u Seznam.cz s velkou pravděpodobností nezáleží. Jako doplněk by mohlo být zajímavé udělat korelaci pozic ve vyhledávacích u domén s exotickými TLD (.travel, .aero atd.).

Vysoká korelace faktorů týkajících se klíčového slova v doméně mě nepřekvapila, viz čtvrtá hypotéza v kapitole 4.4. Stejně tak mě nepřekvapila téměř nulová korelace mezi počtem výskytů klíčového slova na stránce a pozicí stránky na dané slovo v SERP.

Zajímavým zjištěním pro mě byl fakt, že umístění klíčového slova v hlavním nadpisu stránky s výslednou pozicí v SERP koreluje více, než umístění klíčového slova v titulku stránky. Rozdíly v korelačních koeficientech jsou ale velice malé, až zanedbatelné. Přesto jsem čekal, že faktory spojené s umístěním klíčového slova v titulku stránky si budou vést lépe.

Mimo hlavní faktory jsem spočítal korelaci i pro faktory doplňkové:



Obrázek 23: Graf doplňkových faktorů korelační analýzy (zdroj: autor)

Téměř nulová korelace mezi počtem odkazů/odkazujících unikátních webů a pozicí v SERP se také dala očekávat. Vyhledávače hodnotí spíše kvalitu odkazů než jejich absolutní kvantitu. Vhodnější by bylo zahrnout pouze **odkazy obsahující v textu konkrétní klíčové slovo**. Tato data se mi bohužel nepodařilo získat, viz důvody popsané v kapitole 4.1.

Korelační koeficient faktoru klíčového slova v hostname má podobnou hodnotu, jako koeficient faktoru klíčového slova v doméně druhého řádu (viz **Obrázek 22**). Shodný korelační koeficient faktorů *Počet výskytů klíčového slova v hostname* a *Klíčové slovo je obsaženo v hostname* je způsoben faktem, že v naprosté většině případů je klíčové slovo v hostname umístěnou pouze jednou.

Zbývající koeficienty doplňkových faktorů nabývají velmi podobných hodnot, jako příslušné hlavní faktory, viz **Obrázek 23**.

Výpočet průměrného korelačního koeficientu pro všech 17 SEO faktorů zabralo serveru asi **2 minuty**.

4.8. Vyhodnocení hypotéz

1) Nejvyššího korelačního koeficientu dosáhne některý ze SEO faktorů, které popisují umístění klíčového slova v tagu <title> stránky

Tato hypotéza se nepotvrdila. Nejvyššího korelačního koeficientu dosáhl faktor *Klíčové slovo použito v názvu domény* (např. *www.klicoveslovo.cz*) (0,433). Vyššího korelačního koeficientu než faktory popisující umístění klíčového slova v tagu <title> stránky dosáhly překvapivě i faktory, popisující umístění klíčového slova v hlavním nadpisu stránky <h1> (0,27 vs. 0,25).

2) Faktor *Počet výskytů klíčového slova na stránce* se umístí mezi 3 faktory s nejnižším korelačním koeficientem

Hypotéza se potvrdila. Popisovaný faktor dosáhl nejnižší, dokonce lehce záporné, míry korelace -0,051. Z toho vyplývá, že v analyzované sadě dat nebyla nalezena žádná pozitivní souvislost (a jen statisticky nevýznamná negativní souvislost) mezi počtem výskytů klíčového slova na stránce a pozicí dané stránky v SERP.

3) Faktory popisující umístění klíčového slova v tagu <title> budou mít vyšší korelační koeficient než faktory popisující umístění klíčového slova v hlavním nadpisu stránky <h1>

Hypotéza se nepotvrdila. Faktor *Klíčové slovo použito jako první slovo/a v tagu nadpisu <h1>* dosáhl nejvyšší míry korelace 0,27. Naproti tomu faktor *Klíčové slovo použito kdekoliv v tagu <title>* dosáhla podobného výsledku 0,25.

Dá se tedy říci, že umístění klíčového slova do titulku i hlavního nadpisu stránky koreluje s pozicí stránky v SERP velmi podobně.

4) Některý z faktorů, který popisuje umístění klíčového slova v doméně nebo hostname, bude mezi třemi faktory s nejvyšším korelačním koeficientem

Hypotéza se potvrdila. Nejvyššího korelačního koeficientu ze všech analyzovaných faktorů dosáhl faktor *Klíčové slovo použito v názvu domény* (např. *www.klicoveslovo.cz*) (0,433)

4.9. Kompletní porovnání výsledků s odhady důležitosti serveru SEOfactory.cz

Tabulka 13: Srovnání jednotlivých SEO faktorů na základě vypočítaného korelačního koeficientu a procentuální důležitosti ze SEOfactory.cz. (zdroj: autor)

Název faktoru	Korelační koeficient	Důležitost dle SEOfactory.cz
Klíčové slovo použito v názvu domény (např. www.klicoveslovo.cz)	0,433	50 %
Stránka má jinou než českou doménu	0,362	34 %
Stránka má českou (.cz) doménu	0,31	
Klíčové slovo použito jako první slovo/a v tagu nadpisu <h1>	0,27	51 %
Klíčové slovo použito kdekoliv v tagu <title>	0,25	79 %
Klíčové slovo použito kdekoli v tagu nadpisu <h1>	0,216	53 %
Klíčové slovo použito jako první slovo v tagu <title>	0,169	71 %
Klíčové slovo použito v prvních 50 - 100 slovech v HTML kódu stránky	0,044	34 %
Klíčové slovo použito kdekoli v tagu dalších nadpisů <h2> - <h6>	-0,027	44 %
Použití klíčových slov / množství opakování klíčových slov v HTML stránky	-0,051	39 %

Z uvedené tabulky samozřejmě **nelze vydedukovat, že se hodnotitelé mýlili** (vysoká míra korelace neznamená, že daný SEO faktor skutečně přímo ovlivňuje pozici ve vyhledávači).

4.10. Domény, které se nejčastěji objevovaly v analyzovaných výsledcích

Tabulka 14: Seznam domén, které se nejčastěji objevovaly v analyzovaných výsledcích. (zdroj: autor)

#	doména	počet výskytů	#	doména	počet výskytů
1	cs.wikipedia.org	2 308	26	samoleceni.cz	162
2	firmy.cz	1 665	27	turistika.cz	161
3	mimibazar.cz	794	28	facebook.com	161
4	lekarna.cz	695	29	podnikatel.cz	159
5	mall.cz	419	30	edb.cz	158
6	penize.cz	419	31	ordinace.cz	153
7	docsimon.cz	370	32	lekari-online.cz	149
8	mesec.cz	359	33	youtube.com	148
9	rodina.cz	332	34	datart.cz	146
10	alza.cz	319	35	levneelektro.cz	146
11	emimino.cz	316	36	spechato.cz	138
12	en.wikipedia.org	285	37	shopiq.cz	135
13	najisto.centrum.cz	281	38	outletexpert.cz	134
14	muj-bazar.cz	256	39	lekynainternetu.cz	131
15	nakupka.cz	252	40	csas.cz	131
16	newtravel.cz	246	41	lan-shop.cz	129
17	etravel.cz	241	42	auto.bazos.cz	122
18	ulekare.cz	228	43	blackcomb.cz	121
19	omlazeni.cz	221	44	mmreality.cz	121
20	drmax-lekarna.cz	219	45	nabytek-jeko.cz	117
21	gigalekarna.cz	211	46	prozdravi.cz	115
22	finance.cz	203	47	kurzy.cz	115
23	csfd.cz	200	48	cars.cz	113
24	diskuse.doktorka.cz	193	49	otto-shop.cz	111
25	tipcars.com	163	50	skateshop.cz	110

Seznamu webů jasně vévodí Wikipedia, kterou Seznam.cz zobrazil u 1 % klíčových slov v prvních 20 výsledcích vyhledávání. Pokud chtějí firmy získat vyšší návštěvnost

z vyhledávačů, měly by určitě analyzovat možnosti využití stránek Wikipedie, které se umisťují na vysokých pozicích.

Katalog firem firmy.cz se umístil na druhém místě, což ukazuje, že pro firmy má registrace do něj stále smysl včetně testování placeného zvýraznění zápisů pro kategorie, které se umisťují na vysokých pozicích.

4.11. Hlavní problémy, které jsem při zpracování práce musel řešit

Po prvním zpracování dat a výpočtu korelace jsem si uvědomil nutnost mnohem větší normalizace vstupních dat (především odstranění všech bílých znaků z textu mimo jednoduché mezery). Bez této normalizace bylo téměř nemožné napsat správně algoritmy na výpočet hustoty nebo ověření existence víceslovných frází v textu. Mimo normalizace bylo také nutné rozšířit počet jednotkových testů komponent pro kontrolu umístění klíčových slov v doméně, titulku a nadpisu stránek.

Po částečně dokončeném druhém zpracování dat jsem zjistil chybu v SQL dotazu, který vybíral klíčová slova pro analýzu. Před dokončením třetího zpracování dat jsem proto ještě jednou provedl důkladnou kontrolu vstupů a výstupů všech funkcí.

Správného zpracování dat jsem dosáhl především za pomoci:

- jednotkových testů s vysokým pokrytím kódu (většinou 100 %),
- správné dekompozice objektů/problémů a využívání zásady jedné zodpovědnosti tříd²⁰,
- principů Dependency Injection.

Složitým úkolem bylo také převedení obsahu analyzovaných stránek do UTF-8. Narazil jsem především na následující problémy:

- serverové hlavičky ani HTML značka <meta> stránky nedefinovaly kódování stránky v sekci Content-Type,

²⁰ single responsibility principle

- kódování stránky bylo definováno nesmyslně (např. uft-8 nebo, window-1250),
- stránky obsahovaly znaky, které neodpovídaly definovanému kódování a nedaly se proto správně převést do UTF-8.

5. Závěr

5.1. Naplnění cílů

V předchozích kapitolách jsem popisoval způsob, jakým jsem naplnil cíle definované v úvodu praktické části této práce (kapitola 4.1). Konkrétně *„Spočítat korelace mezi jednotlivými SEO faktory, které ovlivňují pozice v SERP na Seznam.cz a skutečnými pozicemi ve výsledcích hledání Seznamu“* a *„rozhodnout, které SEO faktory jsou pro dosažení dobrých pozic pravděpodobně důležitější a které méně důležité“*.

Výsledný okomentovaný graf korelačních koeficientů, sumarizující výsledek mého snažení, je obsažen v **kapitole 4.7.4**. Vyhodnocení hypotéz v **kapitole 4.8**.

Nyní na závěr celé práce jsem si již plně uvědomil, že z výsledků korelační analýzy není ve skutečnosti možné smysluplně rozhodnout, které SEO faktory jsou pro dosažení dobrých pozic důležité. Korelace jednoduše není kauzalita a nelze z ní odvozovat příčiny a následky. V **kapitole 4.9** jsem proto alespoň porovnal pořadí důležitosti SEO faktorů ze zadání s výsledky korelační analýzy.

Na začátku vypracovávání této práce jsem metodu korelační analýzy téměř neznal. Jsem velice rád, že mi diplomová práce pomohla pochopit její možnosti i omezení.

5.2. Možnosti rozšíření práce

Jak jsem psal již v úvodu praktické části, v diplomové práci chybí korelační analýza off-page SEO faktorů analyzovaných stránek. Pokud by se mi někdy v budoucnu podařilo získat podrobná data o zpětných odkazech analyzovaných stránek, bude jednoduché korelační analýzu o off-page SEO faktory rozšířit.

Otázkou je, do jaké míry by byla analýza off-page faktorů (především zpětných odkazů) vypovídající. Vyhledávač Seznam.cz indexuje pouze odkazy ze stránek psaných česky²¹.

²¹ „Pro fulltextové vyhledávání v českém jazyce využívá společnost Seznam.cz, a.s. vlastní technologie, pro hledání v ostatních jazycích (ve světě) je použita technologie Bing.com od firmy Microsoft.“ (28)

Databáze zpětných odkazů popsané v kapitole 3.1 ale jazyk odkazující stránky nerozlišují. Proto mohou být jimi indexované odkazy výrazně odlišné.

Zajímavá by mohla být také analýza, do jaké míry využívá vyhledávač Seznam.cz pro hodnocení stránky v poslední době stále populárnější sociální signály a metriky. Mezi ty patří např. počet sdílení/likes konkrétní stránky na Facebooku, počet tweetů na sociální síti Twitter a další.

6. Terminologický slovník

Termín	Zkratka	Význam (zdroj)
Off-page SEO faktory		<i>„Off page faktory – faktory, které je možné ovlivnit mimo web, patří mezi ně hlavně počet a kvalita zpětných odkazů.“ (6, str. 51)</i>
On-page SEO faktory		<i>„Faktory, které je možné ovlivnit přímo na webu. Patří mezi ně hlavně technický stav webu, struktura a obsah.“ (6, str. 51)</i>
Odchozí odkaz		Odkaz, vedoucí z našeho webu na libovolný jiný web. (vlastní definice autora)
Odkazující stránka		Stránka, na které je umístěn odchozí odkaz. (vlastní definice autora)
Odkazovaná stránka		Stránka, na kterou směřuje odchozí odkaz. (vlastní definice autora)
Míra prokliku (click through rate)	CTR	Počet uživatelů, kteří ve vyhledávači klikli na konkrétní výsledek / počet všech uživatelů, kterým se daný výsledek zobrazil. (vlastní definice autora)
Hostname		Část URL za úvodním <i>http://</i> a před prvním lomítkem. U http://www.seznam.cz/katalog je hostname www.seznam.cz . Někdy se také nazývá jen „doména“. (vlastní definice autora)
Stránka výsledků vyhledávání	SERP	Stránka s výsledky vyhledávání. Tato stránka u hlavních vyhledávačů (Seznam.cz, Google.cz) zobrazuje vždy 10 výsledků s titulkem a krátkým popiskem. (34)
Site-wide odkaz		Odkaz, který je ve stejné podobě umístěn na všech stránkách konkrétního webu (např. dole a stránce v patičce). (vlastní definice autora)

7. Použité zdroje

1. RAND FISHKIN. Keyword Metrics and Data-Driven Actions for SEO - Whiteboard Friday. [online]. 8. srpen 2012 [vid. 22. říjen 2012]. Dostupné z: http://www.youtube.com/watch?v=14uqilubNXg&feature=youtube_gdata_player
2. IAN LURIE. Analytics-Driven SEO: A lesson in 4 steps. [online]. 12. listopad 2009 [vid. 22. říjen 2012]. Dostupné z: <http://www.portent.com/blog/analytics/analytics-driven-seo.htm>
3. MAREK PROKOP. Optimalizace stránek pro vyhledávací a indexovací služby. [online]. 20. srpen 2001 [vid. 1. říjen 2012]. Dostupné z: <http://interval.cz/clanky/optimalizace-stranek-pro-vyhledavaci-a-indexovaci-sluzby/>
4. MICHAL KUBÍČEK. *Velký průvodce SEO: jak dosáhnout nejlepších pozic ve vyhledávačích*. Brno: Computer Press, 2008. ISBN 978-80-251-2195-5.
5. MICHAL KUBÍČEK. *50 způsobů, jak získat zpětný odkaz*. S.l.: PRONETmedia, 2012. ISBN 978-80-87721-01-8.
6. LINDA HLAVÁČOVÁ. *Analýza odkazového profilu* [online]. 2012. S.l.: Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky. Dostupné z: http://library.vse.cz/F/8SALAXQG7CX4YVNUUVS6671QBKMQL4CYLL8HJ9G56SBTNT888H-57462?func=service&doc_library=UEP01&doc_number=000222548&line_number=0002&func_code=WEB-FULL&service_type=MEDIA.
7. ENGE, Eric, Stephan SPENCER, Jessie STRICCHIOLA a Rand FISHKIN. *The Art of SEO*. Second Edition. S.l.: O'Reilly Media, 2012. ISBN 1449304214.
8. SEOMOZ. SEO: The Free Beginner's Guide From SEOmoz. [online]. 2011 [vid. 25. září 2012]. Dostupné z: <http://www.seomoz.org/beginners-guide-to-seo>
9. MAREK PROKOP. Marek Prokop: SEO není optimalizace pro vyhledávače. [online]. 11. březen 2012 [vid. 14. leden 2012]. Dostupné z: <http://www.zdrojak.cz/clanky/marek-prokop-seo-neni-optimalizace-pro-vyhledavace/>
10. ROBERT NĚMEC. Cílem SEO není růst návštěvnosti, cílem SEO je růst zisku. [online]. 29. únor 2012 [vid. 25. září 2012]. Dostupné z: <http://seo-sem.robertnemec.com/cil-seo-navstevnost-zisk/>
11. DUŠAN JANOVSKEÝ. O co jde v SEO -- návštěvnost. [online]. [vid. 25. září 2012]. Dostupné z: <http://www.jakpsatweb.cz/seo/smysl-seo.html>
12. SYMBIO. Long tail. [online]. [vid. 25. září 2012]. Dostupné z: <http://www.symbio.cz/slovník/long-tail.html>
13. MAREK PROKOP. Budování zpětných odkazů (linkbuilding). In: [online]. S.l. 2012. [vid. 27. září 2012]. Dostupné z: https://docs.google.com/present/view?id=dg27v6rj_1056gg5fz9f4.

14. MATT CUTTS. Quick comment on nofollow. [online]. 15. květen 2006 [vid. 30. září 2012]. Dostupné z: <http://www.mattcutts.com/blog/quick-comment-on-nofollow/>
15. BRANKO RIHTMAN. Backlink Data – Who is the Best Provider of them All? [online]. 10. leden 2012 [vid. 12. únor 2012]. Dostupné z: <http://www.rankabove.com/news/seo-research/backlink-data-provider-all/>
16. MAJESTICSEO. Glossary - Majestic SEO. [online]. [vid. 27. leden 2012]. Dostupné z: <http://www.majesticseo.com/support/glossary>
17. MAJESTICSEO. Majestic SEO : Backlink Checker & Site Explorer. [online]. [vid. 12. únor 2012]. Dostupné z: <http://www.majesticseo.com/>
18. SEOMOZ. Open Site Explorer. [online]. [vid. 12. únor 2012]. Dostupné z: <http://www.opensiteexplorer.org/>
19. AHREFS. Ahrefs Site Explorer & Keyword Tool. [online]. 12. leden 2012 [vid. 12. leden 2012]. Dostupné z: <http://ahrefs.com/>
20. BRANDON FRASER. Linkscape Index Delay Explained. [online]. 6. duben 2012 [vid. 26. říjen 2012]. Dostupné z: <http://www.seomoz.org/blog/linkscape-index-delay-explained>
21. DIXON JONES. 7 Ways to get a Free Trial of Majestic SEO. *Majestic SEO Blog* [online]. 29. srpen 2011 [vid. 25. říjen 2012]. Dostupné z: <http://blog.majesticseo.com/general/7-ways-to-get-a-free-trial-of-majestic-seo/>
22. SEOMOZ. Open Site Explorer: hanibal.cz. [online]. [vid. 30. leden 2012]. Dostupné z: http://www.opensiteexplorer.org/anchors?page=1&site=www.hanibal.cz&sort=domains_linking_page&source=phrase&target=domain
23. GOOGLE. Smluvní podmínky společnosti Google – Zásady a pravidla. [online]. 2012 [vid. 29. září 2012]. Dostupné z: <http://www.google.com/intl/cs/policies/terms/>
24. IAN LURIE. How to: Scrape search engines without pissing them off. [online]. 28. září 2011 [vid. 29. září 2012]. Dostupné z: <http://searchnewscentral.com/20110928186/General-SEO/how-to-scrape-search-engines-without-pissing-them-off.html>
25. RAND FISHKIN. 2011 Search Engine Ranking Factors. [online]. 2011 [vid. 13. srpen 2012]. Dostupné z: <http://www.seomoz.org/article/search-ranking-factors#methodology>
26. JAN LINHART. SEO Faktory. *SEO faktory podle důležitosti* [online]. 2012 [vid. 13. srpen 2012]. Dostupné z: <http://www.seofactory.cz/seofactory>
27. PAVEL UNGR. Google uvedl Social Search, další krok k personalizaci výsledků vyhledávání. [online]. 14. leden 2012 [vid. 16. září 2012]. Dostupné z: <http://www.mediar.cz/google-uvedl-social-search-dalsi-krok-k-personalizaci-vysledku-vyhledavani/>

28. SEZNAM.CZ. Zadávání dotazů. [online]. [vid. 16. září 2012]. Dostupné z: <http://napoveda.seznam.cz/cz/zadavani-dotazu.html>
29. VÁCLAV KOHOUT. Spearmanův korelační koeficient. [online]. 18. duben 2012 [vid. 29. září 2012]. Dostupné z: <http://athena.zcu.cz/kurzy/spne/000/HTML/41/>
30. ADAM LUND. php - How to assign a rank number to an array when ties exist. [online]. 11. srpen 2010 [vid. 29. září 2012]. Dostupné z: <http://stackoverflow.com/questions/3463732/how-to-assign-a-rank-number-to-an-array-when-ties-exist>
31. BARCELONA FIELD STUDIES CENTRE S.L. Spearman's Rank Correlation Coefficient. [online]. 26. únor 2011 [vid. 29. září 2012]. Dostupné z: <http://geographyfieldwork.com/SpearmanRank.htm>
32. THE PHP GROUP. PHP: Floating point numbers - Manual. [online]. [vid. 14. srpen 2012]. Dostupné z: <http://php.net/manual/en/language.types.float.php>
33. THE PHP GROUP. PHP: Mathematical Extensions - Manual. [online]. 10. srpen 2012 [vid. 15. srpen 2012]. Dostupné z: <http://www.php.net/manual/en/refs.math.php>
34. SEO - PROFESIONÁL.CZ. SERP. [online]. 2012 [vid. 6. listopad 2012]. Dostupné z: <http://www.seo-profesional.cz/serp/>

8. Seznam obrázků

Obrázek 1: Červená část grafu obsahuje malý počet velmi často hledaných klíčových slov. Naproti tomu modrá část obsahuje velký počet málo hledaných slov (long-tail) (12)	9
Obrázek 2: Rozložení textů odkazů (anchor text) v odkazech webu topgal.cz dle MajesticSEO. (zdroj: autor)	13
Obrázek 3: Poměr dofollow a nofollow odkazů webu Mall.cz dle služby Ahrefs. (zdroj: autor) ..	14
Obrázek 4: Relativní počet zpětných odkazů jak je odhalily zvolené odkazové databáze. Nástroj, který našel nejvíce odkazů, na grafu vždy představuje 100%. (zdroj: autor).....	18
Obrázek 5: Relativní počet nalezených odkazujících domén jak je odhalily zvolené odkazové databáze. Nástroj, který našel nejvíce odkazů, na grafu vždy představuje 100%. (zdroj: autor) ..	19
Obrázek 6: Ukázka rozhraní nástroje MajesticSEO Site Explorer. (zdroj: autor)	20
Obrázek 7: Ukázka rozhraní nástroje Ahrefs Site Explorer. (zdroj: autor).....	21
Obrázek 8: Vývoj celkového počtu odkazů webu hanibal.cz v čase pomocí MajesticSEO. (zdroj: autor)	27
Obrázek 9: Vývoj počtu unikátních domén (viz začátek kapitoly 3.1), které odkazují na web hanibal.cz v čase s pomocí MajesticSEO. (zdroj: autor)	27
Obrázek 10: Vývoj celkového počtu odkazů webu hanibal.cz v čase pomocí MajesticSEO. (zdroj: autor)	28
Obrázek 11: Vývoj počtu unikátních domén, které odkazují na web hanibal.cz v čase s pomocí MajesticSEO. (zdroj: autor).....	28
Obrázek 12: Sekce „Brusle“ e-shopu Sportobchod.cz. Žádný filtr značky není vybrán. (zdroj: autor)	36
Obrázek 13: Sekce „Brusle“ e-shopu Sportobchod.cz. V sekci je zobrazen pouze podvýběr bruslí značky Bauer. (zdroj: autor)	36
Obrázek 14: Výsledky vyhledávání na klíčové slovo flight tickets. Hledaný web kayak.com má ve výsledcích na toto klíčové slovo třetí pozici na google.com. (zdroj: autor).....	40
Obrázek 15: HTML kód prvních tří výsledků Seznam.cz, který parsujeme pomocí regulárních výrazů. (zdroj: autor)	42
Obrázek 16: První tři výsledky při stránkování po 10 výsledcích. (zdroj: autor).....	43
Obrázek 17: První tři výsledky při stránkování po 60 výsledcích. (zdroj: autor).....	43

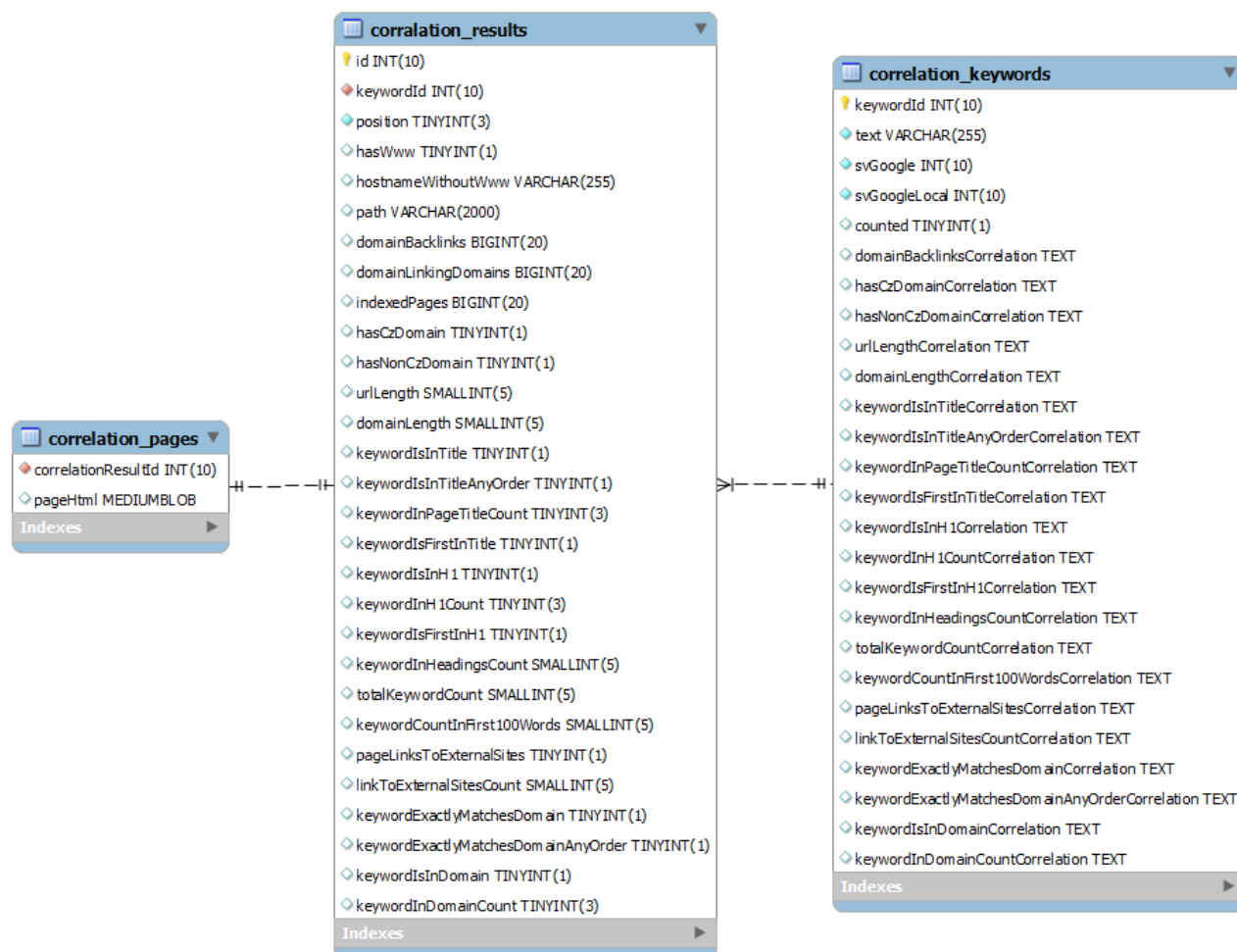
Obrázek 18: Hlášení, které Google zobrazí při odepření přístupu k vyhledávání. (zdroj: autor)	45
Obrázek 19: Schéma detekce problémů vyhledávače při měření pozic. (zdroj: autor)	47
Obrázek 20: Schéma získávání výsledků z vyhledávačů Google a Seznam. (zdroj: autor)	56
Obrázek 21: Vzorec pro výpočet Spearmanova koeficientu pořadové korelace. (29)	61
Obrázek 22: Graf vypočítaných korelačních koeficientů pro definované SEO faktory. (zdroj: autor)	64
Obrázek 23: Graf doplňkových faktorů korelační analýzy (zdroj: autor)	65

9. Seznam tabulek

Tabulka 1: Seznam data-driven SEO nástrojů použitých v práci. (zdroj: autor)	10
Tabulka 2: Informace o velikosti indexu služby Majestic SEO. (17)	16
Tabulka 3: Informace o velikosti indexu služby Open Site Explorer/Linkscape. (18)	17
Tabulka 4: Absolutní počty nalezených zpětných odkazů porovnávaných webů. (zdroj: autor)	18
Tabulka 5: Absolutní počty nalezených odkazujících domén porovnávaných webů. (zdroj: autor)	19
Tabulka 6: Srovnání cen za využívání API pro Případ užití 1. (zdroj: autor)	24
Tabulka 7: Srovnání cen za využívání API pro Případ užití 2. (zdroj: autor)	25
Tabulka 8: Počty nalezených klíčových slov s využitím popsaných nástrojů. (zdroj: autor)	32
Tabulka 9: Výsledná data o hledanosti vybraných klíčových slov. (zdroj: autor)	38
Tabulka 10: Hledanost je vhodné pro lepší přehlednost seskupit podle značky letecké společnosti. (zdroj: autor)	39
Tabulka 11: Distribuce hledanosti vybraných klíčových slov. (zdroj: autor)	56
Tabulka 12: Přehled ohodnocených výsledků vyhledávání na klíčové slovo „domácí pekárny“. (zdroj: autor)	62
Tabulka 13: Srovnání jednotlivých SEO faktorů na základě vypočítaného korelačního koeficientu a procentuální důležitosti ze SEOFaktory.cz. (zdroj: autor)	67
Tabulka 14: Seznam domén, které se nejčastěji objevovaly v analyzovaných výsledcích. (zdroj: autor)	68

10. Přílohy

Příloha 1: Schéma databáze použité pro korelační analýzu



Příloha 2: Předložky a spojky, které byly při normalizaci odstraněny z textu analyzované stránky

od, z, do, bez, krom, kromě, místo, podle, podél, kolem, okolo, u, vedle, během, pomocí, stran, prostřednictvím, k, proti, naproti, oproti, kvůli, díky, pro, za, před, mimo, na, pod, pode, nad, nade, mezi, krom, kromě, skrz, skrze, o, po, v, o, na, v, po, při, s, za, před, pod, nad, mezi, a, i, ani, přímo, tak, hned, jednak, zčásti, dílem, ale, avšak, však, leč, naopak, jenomže, jenže, sice, jistě, ba, nadto, dokonce, nejen, nýbrž, nebo, anebo, buď, totiž,

vždyť, neboť, proto, tudíž, tedy, aby, jakmile, až, než, nežli, zatímco, když, kdyby, pokud, protože, poněvadž

Příloha 3: Algoritmus pro výpočet pořadového ranku vstupů Spearmanovy korelace

```

function countRanks(array $items) {
    $sortedItems = $items;
    rsort($sortedItems);
    $itemCount = count($sortedItems);
    $ranks = array();
    $ranksSum = null;
    $ranksCount = null;

    for ($i = 0; $i < $itemCount; $i++) {
        $itemValue = $sortedItems[$i];
        $itemRankValue = $i + 1;

        if (array_key_exists($i + 1, $sortedItems)) {
            $nextItemValue = $sortedItems[$i + 1];
        }
        else {
            $nextItemValue = null;
        }
        if ($itemValue !== $nextItemValue) {
            if ($ranksSum !== null) {
                $averageRank = ($ranksSum + $itemRankValue) /
                    ($ranksCount + 1);

                $ranks[$itemValue . ''] = $averageRank;
                $ranksSum = null;
                $ranksCount = null;
            }
            else {
                $ranks[$itemValue . ''] = $itemRankValue;
            }
        }
        else {
            $ranksSum += $itemRankValue;
            $ranksCount++;
        }
    }
    $output = array();

    for ($i = 0; $i < $itemCount; $i++) {
        $output[$i] = $ranks[$items[$i] . ''];
    }
    return $output;
}

```