autor: Jiří Koutný, koutny@collabim.cz

Metodika korelační analýzy výsledků vyhledávače Seznam.cz

1.1. Výběr atributů (SEO faktorů) korelace

Faktory jsem vybíral podle průzkumu zveřejněném na webu SEOfaktory.cz. 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. 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.

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.

1.2. 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 1.2.4):

- Klíčové slovo se shoduje s doménou druhého řádu (volná shoda),
- Klíčové slovo je obsaženo v hostname (= celá doména, např. email.seznam.cz),
- 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.

1.2.1. 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>	
Upřesněný název:	Klíčové slovo použito kdekoliv v tagu <title> (přesná shoda)
Popis:	**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.
Důležitost:	SEOFaktory.cz: **79 %**

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

1.2.2. 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>	
Upřesněný název:	Klíčové slovo je obsaženo v prvním tagu H1
Popis:	Z toho faktoru jsem odvodil doplňkový faktor **Počet výskytů klíčového slova v prvním tagu H1**.
Důležitost:	SEOFaktory.cz: **53 %**

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

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

1.2.3. Atributy ostatního obsahu stránky

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

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

1.2.4. Atributy domény a URL

Klíčové slovo použito v názvu domény (např. www.klicoveslovo.cz)	
Upřesněný název:	Klíčové slovo se shoduje s doménou **druhého řádu** (přesná shoda)
Popis:	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")
Důležitost:	SEOFaktory.cz: **50 %**

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

1.3. 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 náhodně vybraná klíčová slova klientů Collabimu.

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

- vybíráme obecnější klíčová slova, která mají hledanost v Česku větší než **250 hledání měsíčně** (hledanost na Google). Číslo 250 jsem vybral tak, aby analyzovaná skupina obsahovala kolem 10 000 klíčových slov (stejně jako v případě analýzy SEOMozu),
- vybíráme pouze klíčová slova, pro která Collabim změřil pozice **nejdéle před jedním dnem**,
- kvůli zjednodušení nás zajímají nás pouze fráze spojené z maximálně **3 slov**,
- opět kvůli zjednodušení vybíráme pouze fráze **bez nealfanumerických znaků** (čárky, apostrofy, uvozovky atd.).

Vstupem pro analýzu je celkem **11 363 klíčových slov**.

***Tabulka 1**: 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

1.4. Zpracování výsledků

K vybraným klíčovým slovům připadá 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.

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.

1.4.1. 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.

1.4.2. Výpočet hodnot všech SEO faktorů pro jednotlivé stránky

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.
- **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 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.
- **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 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 mnohem jednodušeji dále zpracovatelných 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,

1.4.3. 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).

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

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“).

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

Spearmanův korelační koeficient jsem spočítal **pro každé klíčové slovo zvlášť** (tedy vždy z 20 výsledků pro daný SEO faktor).

1.5. Možnosti rozšíření práce

V 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 česky.

Databáze zpětných odkazů jako SEOMoz, MajesticSEO nebo Ahrefs ale jazyk odkazující stránky nerozlišují, a proto mohou být odkazy 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.

Zaujala vás má analýza? Našli jste v ní další zajímavé vztahy, které jsem přehlédl? Budu rád, když mi **napíšete na koutny@collabim.cz**.