

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

INFORMAČNÍ SYSTÉM PRO ANALÝZU SEO PARAMETRŮ WEBOVÝCH PREZENTACÍ

BAKALÁŘSKÁ PRÁCE

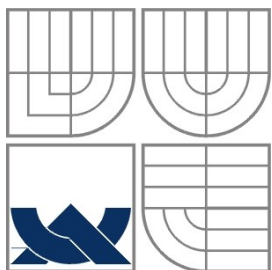
BACHELOR'S THESIS

AUTOR PRÁCE

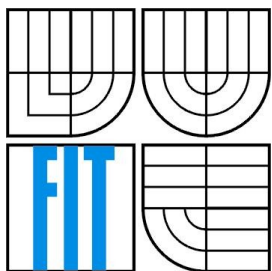
AUTHOR

ONDŘEJ KARLÍK

BRNO 2009



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

INFORMAČNÍ SYSTÉM PRO ANALÝZU SEO PARAMETRŮ WEBOVÝCH PREZENTACÍ

SEARCH ENGINE OPTIMIZATION (SEO) ANALYSIS INFORMATION SYSTEM

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

ONDŘEJ KARLÍK

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. JIŘÍ TOBOLA

BRNO 2009

Abstrakt

Tato práce pojednává o problematice zpřístupnění často využívaných informací uživatelům provádějícím optimalizace webových stránek pro vyhledávače implementovaným informačním systémem. Obsahuje teoretický popis sledovaných SEO parametrů a způsobů jejich získání a zpracování, také vysvětlení úskalí širšího využití informačního systému.

Klíčová slova

optimalizace webových stránek pro vyhledávače, SEO, OLAP, distribuované zpracování, protokol, PHP, MySQL, HTML, CSS, jQuery

Abstract

This thesis describes collecting of recently used information for users optimizing their web applications for search engines. Details of SEO parameters and principles of their processing are discussed here. The thesis also learns about problems appearing when the information system is used frequently.

Keywords

search engine optimization, SEO, OLAP, distributed processing, protocol, PHP, MySQL, HTML, CSS, jQuery

Citace

Ondřej Karlík: Informační systém pro analýzu SEO parametrů webových prezentací, bakalářská práce, Brno, FIT VUT v Brně, 2009

Informační systém pro analýzu SEO parametrů webových prezentací

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením Ing. Jiřího Toboly.

Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Ondřej Karlík
12. května 2009

Poděkování

Rád bych poděkoval vedoucímu práce Ing. Tobolovi za vstřícný přístup a Ing. Kurkovi, projektovému manažerovi Via Aurea s.r.o., za poskytnutí konzultací, připomínek a možnosti využití serverů společnosti Via Aurea pro testování aplikace.

© Ondřej Karlík, 2009.

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

Obsah

Obsah.....	1
1 Úvod.....	2
2 SEO.....	3
2.1 Vyhledávače.....	3
2.1.1 Obecný popis funkcí vyhledávačů.....	4
2.1.2 Nejdůležitější představitelé na českém trhu.....	4
2.2 Klíčová slova.....	5
2.3 Zpětné odkazy.....	7
2.4 Validita HTML kódu.....	7
2.5 Další parametry.....	8
3 Informační systém.....	9
3.1 Obecný popis systému.....	9
3.1.1 Uživatelské požadavky.....	9
3.2 Detaily implementace.....	10
3.2.1 Technické požadavky.....	10
3.2.2 Použité knihovny.....	10
3.2.3 Framework.....	11
3.2.4 Uživatelské rozhraní.....	12
3.2.5 Bezpečnost.....	13
3.2.6 Oprávnění.....	15
3.3 Získávání dat.....	16
3.3.1 Problémy masového nasazení systému.....	17
3.3.2 Architektura a protokol.....	18
3.3.3 Výhody a nevýhody zvoleného řešení.....	20
3.4 Zpracování dat.....	21
3.4.1 Sumarizace a agregace.....	21
3.4.2 Archivace dat.....	22
3.4.3 Chybějící informace.....	22
3.5 Umístění demonstrační verze systému.....	23
4 Závěr.....	24
Literatura.....	25
Seznam příloh.....	26
Příloha 1: Návod k instalaci systému.....	27
Příloha 2: Vzor postupu při optimalizaci webové stránky.....	28

1 Úvod

Optimalizace webových stránek pro vyhledávače (SEO) je v současnosti velmi důležitým nástrojem pro zvýšení návštěvnosti stránek a tak zejména zvýšení prodeje či prestiže jejich vlastníkům. Tzv. SEO analýzou se v České republice zabývá mnoho odborníků, avšak ti jsou většinou odkázáni na časově náročnou práci hledání způsobů, jak webové stránky ve velké konkurenci protlačit ve výsledcích vyhledávání výše.

Čtenář neznalý základů SEO jistě ocení přílohu této práce, která na fiktivním příkladu ukazuje postup analýzy a úvodní optimalizace. Problematika SEO analýzy bude detailněji vysvětlena v následujících kapitolách na vzoru implementovaného informačního systému, jehož úkolem je rychle a přívětivě podat analytikům všechny potřebné informace včetně přidáných užitečných funkcí. Práce ale nejdříve provede čtenáře shrnutím aktuální situace českého internetového trhu ve vztahu k nabízeným službám analýze nápomocným, aby později právě na nich mohla ukázat schopnosti informačního systému. Dále také popíše podstatu a protokol pro distribuované zpracování, avšak podrobnější technické detaily implementace čtenář najde v projektové dokumentaci.

2 SEO

Základním kamenem optimalizace pro vyhledávače je správná volba klíčových slov, na něž se budou optimalizované webové stránky soustředit. Konkrétně je třeba vzít v úvahu klíčová slova, která nejlépe definují obor podnikání zákazníka, avšak u nichž je dostatečná šance na úspěch. Ta nemusí být vždy jistá - ať už z důvodu malé četnosti vyhledávání těchto slov uživateli, nebo právě pro velkou konkurenci na trhu. Na řadu pak, je-li to možné, přichází úprava stránek - vybraná klíčová slova je vhodné umístit na kritická místa, například do nadpisů, odkazů či přímo do adresy. Pokud pak na analyzované stránky odkazuje řada kvalitních webových stránek (například díky výměně odkazů či inzerci), téměř jistě se dostaví kladný výsledek.

Úvodní popis pravděpodobně ukazuje SEO analýzu jako jednoduchý proces, avšak ve skutečnosti vyžaduje řadu zkušeností, trpělivost a také jistou dávku znalostí z marketingu. Zdrojů potřebných informací je několik, navíc se liší v každé zemi, jednotlivé vyhledávače fungují na jiných principech a konkurenční webové stránky se mohou každý den ve výsledcích vyhledávání dostat na lepší pozici a tím zmíněné prodeje znovu snížit.

V následujících kapitolách bude čtenáři představeno několik základních pilířů SEO analýzy, tzv. *on-page* a *off-page* faktorů, které se jí přímo týkají, a také teorie nezbytná pro pochopení problematiky.

2.1 Vyhledávače

Vyhledávače, zde tedy webové servery poskytující dotazujícím se uživatelům relevantně uspořádané výsledky na jejich hledané výrazy, jsou jedním ze zdrojů návštěvnosti webových stránek. Mezi ty další pak patří přímé přístupy, kdy uživatelé zadávají adresy požadovaných webových stránek přímo do adresního řádku, nebo také vzájemné provázání jednotlivých webových stránek, byť identifikace těchto zdrojů nemusí být jednoznačná - probíhá totiž na základě informace zvané *referrer* (v češtině odkazující stránka) zasílané webovým prohlížečem, kterou lze podvrhnout či její zaslání úplně zakázat.

Vyhledávače jsou však, neuvažujeme-li však placené záznamy služeb například ve formě katalogů firem či reklamních kampaní, zdrojem, který nepotřebuje vstupní investici, a zároveň oslovují velké množství zákazníků. Pro malé společnosti tak může jít o jedinou možnost, jak potencionální klienty oslovit.

2.1.1 Obecný popis funkcí vyhledávačů

Základní částí vyhledávače je tzv. robot (v této terminologii jsou synonymy termíny *bot*, *crawler*, *spider* či *indexer* [1]), což je aplikace procházející webové stránky (také např. dokumenty typu PDF, DOC, RTF a další) a archivuje jejich obsah, přesněji indexuje jej. V tomto obsahu pak po zpracování za pomoci vlastních vyhledávacích algoritmů hledá stránky relevantní dotazům uživatelů. Uvažuje duplicitu obsahu, řeší vztahy mezi stránkami a v neposlední řadě na základě relevance výsledků a jejich kvality a návštěvnosti „známkuje“ webové stránky, jak je vysvětleno v kapitole 2.3 o zpětných odkazech.

Zpracování bere také v potaz váhu jednotlivých částí stránky, protože například sousloví uvedené v nadpisu stránky bude ve většině případů mnohem relevantnější než stejné sousloví v některém odstavci. Váhy budou popsány v kapitole 2.2 věnující se klíčovému slovu.

Pro procházení webu jsou využívány odkazy - pokud na stránku neexistuje žádný odkaz, crawler ji nenajde. Z tohoto důvodu existuje nástroj *sitemap* ve formě XML souboru s odkazy na všechny stránky daného serveru - odkazy pak mohou být do fáze indexování zahrnuty také.

Pro český internetový trh je vhodné zmínit čtveřici serverů poskytujících vyhledávání v Internetu. Další vyhledávače mají zanedbatelné využití, využívají služeb uvedených vyhledávačů nebo zároveň neposkytují žádné funkce pro SEO analýzu užitečné.

2.1.2 Nejdůležitější představitelé na českém trhu

2.1.2.1 Seznam.cz

Seznam.cz byl založen v roce 1996 a od té doby má na trhu z původně českých vyhledávačů největší zastoupení - v červnu 2008 šlo o 62,8% trhu [3]. Je však nutné dodat, že tento server poskytuje cizojazyčné či zahraniční výsledky prostřednictvím dále popsaného serveru firmy Google, přímo se tak zaměřuje jen na české webové stránky. Od poloviny roku 2009 pak bude zahraniční výsledky poskytovat firma Microsoft se službou *Live Search*, čímž na trh vstoupí další důležitý hráč [6].

Server pro inzerenty zpřístupnil statistiky udávající počty hledání jednotlivých výrazů v posledním týdnu a dni společně s častými slovními spojeními. Při sestavování klíčových slov optimalizace je toto jedním z nejdůležitějších ukazatelů. Dále zveřejňuje hodnocení jednotlivých webových stránek, tzv. *S-Rank*, využívané jako ukazatel kvality stránek z pohledu SEO, zejména tedy počtu a kvality zpětných odkazů.

Provozuje služby reklamního programu Sklik či vlastního katalogu firem a zboží s možností vkládání neplacených i placených záznamů.

2.1.2.2 Google.com

Celosvětově nejpoužívanější vyhledávací služba firmy Google měla v ČR v červnu 2008 zastoupení 30% [3]. Česká republika je tak jednou z pěti zemí, kde vyhledávač Google nedrží tržní prvenství [4].

Vzhledem k široké uživatelské základně nabízí pro SEO analýzu několik užitečných nástrojů, například pro zjištění počtu zpětných odkazů či indexovaných stránek pro danou doménu, analýzu slovních spojení včetně ukazatele množství konkurence v reklamním systému (pro vlastní výsledky vyhledávání), ukazatel oblíbenosti *PageRank* nebo například rozsáhlý systém sledování statistik návštěvníků *Google Analytics*.

Provozuje reklamní program *AdWords* a možnost placeného upřednostnění záznamů ve výsledcích vyhledávání.

2.1.2.3 Jyxo.cz

Původní český vyhledávač se v současnosti soustřeďuje na poskytování výsledků vyhledávání ve specializovaných databázích, např. zboží nebo multimédií [5]. Jeho technologie jsou integrovány do řady portálů, samostatně má však tento vyhledávač již minimální návštěvnost [3].

Vyhledávač informuje o tzv. *JyxoRanku* - jde o ukazatel totožný S-Ranku a PageRanku.

2.1.2.4 Yahoo.com

Na zahraničním trhu zabírá Yahoo! druhou příčku v počtu hledaných výrazů, v ČR má minimální zastoupení [7, 3]. Poskytuje mj. nástroje pro zjištění stránek odkazujících na analyzovanou stránku.

2.2 Klíčová slova

Při návrhu klíčových slov pro optimalizované webové stránky může napomoci nástroj *AdWords* zmíněný v předchozí kapitole. Jak ukazuje obrázek na další straně, nástroj uvádí přibližné počty hledání jednotlivých výrazů a množství inzerující konkurence podle těchto výrazů. Samotný výběr slov pak záleží na ochotě riskovat velkou konkurencí, na plánech rozvoje sortimentu firmy v závislosti na situaci na trhu (například orientace na medikamenty v době očekávané pandemie chřipky) či na schopnosti sestavit klíčová slova, která přinesou požadované zvýšení poptávky i mimo hlavní proud konkurence. Pokud ale stránka zohledňuje faktory ovlivňující ukazatele oblíbenosti a kvality, tzv. *ranky*, a odkazuje na ni dostatek kvalitních webů například díky výměně odkazů mezi nimi, může se do konkurenčního boje důležitého trhu pustit ihned.

Samotnými klíčovými slovy pak bývají výrazy uváděné v titulcích stránek, nadpisech, popisech obrázků a odkazů, také ve zbytku obsahu - to v různých tvarech i slovních spojeních, aby pokryly největší možné rozpětí uživateli vyhledávaných variant výrazů.

Keywords	Advertiser Competition ?	Approx Avg Search Volume ?	Search Volume Trends ?
Keywords related to term(s) entered - sorted by relevance ?			
notebooks		2,240,000	
hp notebooks		74,000	
toshiba notebooks		49,500	
cheap notebooks		33,100	
refurbished notebooks		27,100	
acer notebooks		33,100	
asus notebooks		12,100	
laptop notebooks		301,000	
notebooks laptops		301,000	
best notebooks		18,100	
computer notebooks		823,000	
notebooks computers		450,000	
compaq notebooks		12,100	
centrino notebooks		9,900	
samsung notebooks		3,600	
used notebooks		33,100	

Obrázek 2.1: Google AdWords, nápomoc při tvorbě klíčových slov

Důležitost jednotlivých prvků webové stránky z hlediska umístění klíčových slov je různá v rámci interních algoritmů všech vyhledávačů. Obecně uplatňované hodnocení však odpovídá následujícímu seznamu, kde dříve uvedené prvky mají důležitost vyšší [8].

1. titulek stránky <title>
2. text odkazů (zpravidla podtržený text)
3. klíčová slova v textu dokumentu
4. interní odkazy na dokument ze stejného webu
5. externí odkazy na odkazující stránky
6. odkazová popularita webu v tematické skupině
7. celková odkazová popularita webu

Zadá-li potencionální zákazník slovo či sousloví, pro něhož byla stránka optimalizována, v nejlepším případě uvidí tuto stránku na první příčce seznamu s výsledky vyhledávání. Protože však vyhledávač zohledňuje i pořadí slov hledaného výrazu, vzdálenost slov v textu či frekvenci užití těchto slov, je opravdu vhodné klíčová slova uvádět ve více uvažovatelných variantách.

Pořadí optimalizované webové stránky ve výsledcích vyhledávání podle klíčových slov je tedy hlavním kritériem úspěchu volby klíčových slov. Nejjednodušší způsob, jak automatizovaně zjistit toto pořadí, je simulace hledání zajímavých klíčových slov a prohledání vrácené stránky s výsledky hledání, jestli se v prvních desítkách výsledků objeví doména optimalizované webové stránky.

2.3 Zpětné odkazy

Algoritmy vyhledávačů ohodnocují jednotlivé webové stránky, aby mohly sestavit žebříčky kvalitních a důvěryhodných zdrojů. Pokud pak kvalitní zdroj odkazuje na zdroj nový, předpokládá se i jeho kvalita - z tohoto důvodu vznikl systém výměny odkazů, často i jejich nákupu a prodeje.

Uvedený systém s sebou však nutně nese riziko ztráty důvěryhodnosti v případě citování a odkazování tzv. farmami odkazů, jejichž praktiky bývají často nekalé a tím pádem i vyhledávači penalizované.

Váha zpětných odkazů uvedená v předchozí kapitole (položky 2, 5, 6 a 7 seznamu) ukazuje tento parametr jako velmi důležitý. Pro analytika je výhodné sledovat, které weby na optimalizované stránky odkazují, jejich počet zvyšovat a zvyšovat jejich udávanou kvalitu a důvěryhodnost. Problematika výpočtu algoritmů ranků a výměny odkazů přesahuje téma této práce, čtenáři doporučuji uvedené zdroje literatury [2, 9].

Kritériem hodnocení úspěchu optimalizace zpětných odkazů je tedy zvýšení ranků, tedy ukazatelů kvality a důvěryhodnosti stránek. Samotné ranky jednotlivých webových stránek jsou vyhledávači zveřejňovány, například Seznam.cz vrácením výsledku vzdáleného volání procedury ve formátu XML (XML-RPC).

2.4 Validita HTML kódu

Validní kód, tedy kód, který uvádí a následně dodržuje zvolený standardizovaný typ dokumentu udávající správnou syntax, je vizitkou programátora či kodéra. V problematice SEO však validita kódu zaujímá důležitou roli - při zpracovávání webových stránek algoritmy vyhledávačů mohou v některých případech u nevalidní stránky způsobit nesprávné zaindexování obsahu. Příkladem může být chybějící uzavírací značka nadpisu H1, která zapříčiní nechtěné zpracování následující části stránky, od několika slov až po celý odstavec, jako pro dokument klíčové. Algoritmus toto chování navíc může ohodnotit jako nepřipustné a penalizovat jej snížením hodnocení stránky, protože připomíná techniku zneužití tohoto upřednostňovaného prvku stránky.

Uvedená problematika validity kódu se týká jazyka HTML, avšak je nutné dodat, že podobné problémy se zpracováním se mohou vyskytovat i v dalších typech dokumentů. Od XML (častým projevem může být nečitelnost zpráv RSS kanálu) až po CSS, kdy může webový prohlížeč interpretovat stylový předpis nesprávně a v krajních případech zamezit čitelnosti celé webové stránky. Dodržení standardů pak má za úkol zajistit bezproblémové zobrazení informací pro všechny uživatele.

Validní kód se tedy promítá ve správném indexování stránek, lze jej zkontrolovat takzvanými *validátory* - aplikacemi sledujícími, zda-li kód odpovídá definicím typů dokumentů. Nejznámějšími validátory jsou open-source produkty WWW konzorcia (W3C).

2.5 Další parametry

Mezi další parametry SEO analýzy patří například:

- duplicita, kdy více odkazů směřuje na stejný obsah - celkové hodnocení stránek se pak nevýhodně dělí na tyto stránky,
- přepisování URL, tedy technika umožňující uvádění klíčových slov přímo v odkazech, čímž přidává možné umístění klíčových slov a zvyšuje uživatelský komfort,
- překážky pro prohlížeče například v podobě skriptování na straně klientského počítače (JavaScript) či jinak pro vyhledávače sémanticky nelogického či vůbec nezobrazeného kódu - rámce, technologie *Adobe Flash*, data zobrazovaná za pomoci technologie *AJAX*, rozbalovací menu apod.

Obecné rozdělení všech parametrů udávají dvě kategorie. Takzvané *on-page* faktory SEO optimalizace jsou obvykle v přímé režii programátora webových stránek, popřípadě analytika. Patří mezi ně volba klíčových slov a jejich umístění v obsahu či nadpisech, také validita kódu či poslední tři uvedené parametry. Naopak mezi *off-page* faktory se řadí ty, jež vznikají při vytváření struktury ukazatelů kvalitních a důvěryhodných stránek vyhledávači, ať už v rámci tématického seskupení webových stránek či globálních ranků.

Jak již bylo zmíněno, SEO optimalizace ukrývá možnosti zneužití známých nešvarů algoritmů vyhledávačů - příkladem může být tzv. *Google bomb*, která je důsledkem vysokého hodnocení textu odkazů způsobujícího skutečnost, že se poškozená stránka objeví ve výsledcích vyhledávání například vulgárního výrazu, aniž by tento výraz obsahovala. Dále lze uvést techniku skrývání klíčových slov do textu tak, aby je viděl pouze crawler - tato technika umožní zvýšení atraktivnosti podvodných webových stránek za cenu možnosti dlouhodobé penalizace vyhledávačem.

3 Informační systém

3.1 Obecný popis systému

Informační systém již svým oborem působnosti udává úroveň znalostí jeho uživatelů, analytiků. Zároveň přináší, vzhledem k dále popsané problematice distribuovaného zpracování, další netradiční požadavky - synchronizaci času i nutnost spouštění skriptu ve zvoleném časovém intervalu. Tyto i další požadavky práce uvádí v následující kapitole.

3.1.1 Uživatelské požadavky

Uživatelské rozhraní systému se může více soustředit na otázky funkčnosti a použití přesných jednoznačných termínů namísto zachování pochopitelnosti systému laiky. Předpokládá se totiž, že analytici jsou před použitím systému proškoleni a problematiku SEO analýzy znají.

Protože systém pracuje s citlivými informacemi, například s chystanými změnami SEO parametrů či s dosud nezveřejněnými adresami stránek, jež mohou být zneužity v soutěži o zákazníky, důležitým požadavkem je také bezpečnost systému. Krokům podniknutým pro zvýšení bezpečnosti systému se věnuje kapitola 3.2.5.

Obecné požadavky, jako je například vícejazyčnost systému, možnost běhu ve všech obvyklých webových prohlížečích či pravidla přístupnosti a zachování funkčnosti v případě zobrazení v prohlížeči bez nainstalované podpory technologie Adobe Flash či interpretace jazyka JavaScript, jsou samozřejmostí. Systém uvedené požadavky bez výjimky zohledňuje.

Jak vysvětluje kapitola 3.3, separátní aplikace pro získávání dat poběží na několika počítačích zároveň. V případě opravdu masového nasazení systému se pak může očekávat, že všechny servery s touto aplikací nebudou fyzicky umístěny ve stejném časovém pásmu. Zpracování fronty požadavků by se tak mohlo zkomplikovat velkými rozdíly v záznamech provedených operací - tento problém řeší jednoduché nastavení rozdílu oproti zvolenému časovému pásmu (nejlépe tedy tomu, v němž je umístěn HTTP server zobrazující rozhraní informačního systému). Každý počítač, kde aplikace běží, potom ve speciálním konfiguračním souboru ponese informaci o rozdílu času a tuto informaci pak při ukládání záznamů zahrne.

Dalším řešením uvedeného problému by mohlo být nastavení stejného časového pásma na všech serverech bez ohledu na jejich lokalitu, avšak v případě, že by servery nebyly speciálně určené pro běh této aplikace (například pokud by šlo o běžné webové servery poskytující služby webhostingu nezávislým systémům), ostatní aplikace by byly rozdílem ovlivněny také.

Z důvodu distribuovaného zpracovávání fronty požadavků je pro vyšší přesnost také vhodné provádět na počítačích, kde aplikace běží, synchronizaci času protokolem NTP. Problematika časové synchronizace bude vysvětlena znovu v kapitole 3.3 věnující se návrhu protokolu.

3.2 Detaily implementace

Podkapitoly níže stručně provedou čtenáře informacemi o specifikách implementovaného systému z pohledu netypických změn či rozšíření, samotné uživatelské rozhraní a jeho prvky čtenář může, společně s popisem v manuálu, vidět díky veřejné demonstrační verzi systému, případně ze zachycených obrazovek (screenshotů) přiložených na CD.

3.2.1 Technické požadavky

Zdrojové kódy systému byly napsány ve skriptovacím jazyce PHP, pro běh systému je vyžadována minimální verze interpretu 5 s nainstalovaným modulem *cURL*. Databázové prostředí poskytuje *MySQL*, znovu je pro správnou funkčnost vyžadována nejnižší verze 5. Aplikaci pak uživatelům zpřístupňuje webový server *Apache*, za jehož pomoci byl systém vyvíjen. Server pak musí obsahovat modul pro PHP a dále modul *mod_rewrite* - bez jeho podpory bude aplikace zobrazovat po kliknutí na většinu odkazů chybové hlášení o nenalezení souboru.

Pro dosažení vyššího výkonu lze pak použít server *Lighttpd*, na němž byl systém také testován, je však nutné provést několik úprav v konfiguraci, bližší detaily jsou uvedeny v programové dokumentaci.

Systém je díky výše uvedeným aplikacím multiplatformní a staví na otevřenosti zdrojového kódu všech součástí - včetně dále popsanych modulů a knihoven.

3.2.2 Použité knihovny

Pro validaci (syntaktickou a sémantickou kontrolu) HTML kódu sledovaných stránek a jejich stylů systém využívá volně dostupných aplikací organizace W3C s názvem *HTML Validator*, resp. *CSS Validator*. Software na validaci kaskádových stylů vyžaduje, vzhledem ke skutečnosti, že je napsán v jazyce Java, instalaci příslušného prostředí. Grafy jsou v projektu vykreslovány za pomoci knihoven *LibChart* (obrázkové grafy) a *Open Flash Chart* (grafy prostředí Adobe Flash). Pro zjištění geografické lokace podle IP adresy je využito rozhraní *GeoIP* firmy MaxMind.

Všechny knihovny nebo moduly jsou zveřejněny pod svobodnou licencí, zejména *GNU GPL* a *GNU LGPL*. Nachází se ve složce *3rdParty* adresářové struktury frameworku, aby došlo k jejich oddělení od vlastního kódu.

3.2.3 Framework

Znovupoužitelnost informačního systému je obecným požadavkem, jemuž z velké části vyhoví aplikace postavená na dobrém *frameworku*, tedy často od kódu systému oddělené struktuře nižší úrovně, která poskytuje vývojáři využívané funkce nezávislé na výsledném konkrétním použití systému a tím vývoj zrychluje, zpřehledňuje a hlavně vyšší abstrakcí jednotek redukuje počet chyb v celém systému.

Výhodu znovupoužitelnosti doplňuje tzv. *model-view-controller* (MVC), tedy architektura prohlubující abstrakci oddělením interní reprezentace informací, jejich prezentace a samotného řízení chování aplikace [13].

Z uvedených důvodů a také pro potřebu možnosti jednoduché změny vzhledu, jak je vysvětleno v kapitole 3.2.6, jsem pro realizaci systému volil kombinaci obou přístupů. Jádru systému je tvořeno frameworkem, jehož složení nejlépe popíše následující seznam některých částí:

- databázové rozhraní pro intuitivnější a bezpečnější práci se SQL
- třída autentizace s možností automatického blokování přístupu k účtu místo odhlášení
- třída internacionalizace a přidružených funkcí, například skloňování
- rozhraní pro reprezentaci informací v seznamech (vč. grafů)
- dále lze uvést třídy pro práci s řetězci, časem, soubory či grafy

Systém pro separaci grafické prezentace informací v hojné míře využívá šablon, tedy řetězců s předurčenými místy, do nichž budou, třeba i obdobně zpracované, vkládány další podřetězce. Příkladem tedy může být kostra celého webu, do níž bude na místo mezi párem značek `title` vkládán právě titulek stránky v závislosti na adrese stránky.

Samotná řídicí část, tedy *controller*, je vyjádřena třídou zpracovávající celý obsah v kontextu zadaných parametrů jednotlivých modulů - tyto moduly pak nastavují parametry dalším objektům, například seznamům s grafy. Parametry, stejně jako tomu je u dále popsaného získávání dat, jsou pak obyčejnými poli s definovanou strukturou, jejichž výhodou je vysoká univerzálnost. Tyto tedy, společně s databázovým rozhraním, tvoří zbývající část *model*.

Z důvodu nezávislosti na dalším modulu, v tomto případě *GNU gettext*, a také kvůli jednodušší správě jsem implementoval vlastní rozhraní pro internacionalizaci systému. Místo obvyklé funkce `_()` systém všude používá vlastní funkci `_l()` lišící se v několika ohledech. Vlastní funkce využívá globální pole řetězců uložené v jednom ze souborů s překlady výrazů vybraného jazyka. Stejně jako u původní funkce `gettext` je v případě nenalezení řetězce použit řetězec, který je klíčem pro přístup do zmíněného pole, avšak vlastní funkce je schopna generovat, je-li tato možnost v konfiguraci povolena, do souboru seznam těchto nepřeložených výrazů v jednoduše použitelné

formě. Jsou tak vytvářeny soubory, díky nimž je pro překladatele mnohem snazší objevit nepřeložené výrazy a podle nich pak rychle zareagovat pouhou úpravou souboru s překlady.

Důležitým požadavkem na framework je jednoduchost jeho integrace do existujícího systému. Implementovaný framework využívá výhod plynoucích z použití prefixů v názvech tříd (dalším řešením jsou tzv. jmenné prostory, avšak PHP verze 5.3, která je podporuje, v současnosti ještě není ve finální verzi), v tomto případě SEO_, čímž velmi pravděpodobně nedojde ke kolizi s třídami použitými v systému, do něhož framework integrujeme. Uvedený prefix je samozřejmě vzorový, v komerčním využití by spíše obsahoval například zkratku názvu firmy či univerzální označení v podobě názvu celého frameworku. Prefixy jsou využity také v názvech tabulek databáze, kde znovu jednoduše minimalizují riziko kolize.

Při integraci, pomineme-li nedodržení některého z bodů uvedených návodu na instalaci v příloze 1, pak mohou vzniknout konflikty kvůli názvům obecných funkcí, jako je již zmíněná `_l()` či funkce na automatické nahrávání souborů podle názvů volaných a dosud neznámých tříd `__autoload()`.

Uživatelské rozhraní využívá ve velké míře populárního *jQuery*, tedy frameworku pro jazyk JavaScript. Mezi jeho výhody patří také přenositelnost a, v porovnání s ostatními frameworky tohoto druhu, snadná integrovatelnost. Zároveň minimalizuje nekorektní či nekompletní podporu a přidává jednotnost specifických částí jazyka JavaScript v rámci webových prohlížečů, k pak nesčetně nových funkcí zjednodušujících práci s tzv. objektovým modelem dokumentu (DOM) apod.

3.2.4 Uživatelské rozhraní

The screenshot displays the 'SEOanalytics' web application. At the top, a dark header bar contains the logo 'SEOanalytics' on the left and user navigation links on the right: 'přihlášen andřej karlík', 'projekt', and 'volby odhlásit se'. A left sidebar lists navigation options: 'Úvodní strana', 'Ranky', 'Klíčová slova', 'Vyhledávače', 'Odkazy', 'Syntax a sémantika', 'Obecné informace', 'Notifikace', and 'Nastavení'. The main content area is titled 'Zaslání zprávy, hlášení o chybě či nápadu' and includes instructions: 'Následujícím formulářem můžete rychle a snadno odeslat libovolnou zprávu autorovi systému, oznámit mu chybu, připomínku či Váš nápad na vylepšení. Vyberte, prosíme, typ zprávy a vyplňte požadovaná pole.' Below this, a sub-header reads 'Hlášení chyb a nápadů, kontaktní formulář'. The form contains three fields: 'Typ zprávy:' with a dropdown menu set to 'Kontakt', 'Vaše jméno:' with a text input field, and 'Vaše e-mailová adresa:' with a text input field. A large text area for 'Text zprávy:' is also present. An 'Odeslat' button is at the bottom of the form. The footer of the page indicates '© 2008-2009 Ondřej Karlík'.

Obrázek 3.1: příklad obrazovky uživatelského rozhraní, oranžově orámovaná pole jsou povinná

Uživatelské rozhraní klade důraz na přehlednost a jednoduchost při dodržení pravidel přístupnosti a využití inovativních prvků. Systém je plně funkční v prohlížeči bez podpory

- CSS, díky blokům odděleným čarami neztrácí na přehlednosti,
- JavaScriptu, například tlačítko se seznamem voleb či projektů uživatele směřuje pomocí odeslání formuláře s přednastavenou akcí,
- technologie Flash, grafy jsou nahrazeny obrázkovými grafy ve formátu PNG.

Uvedené překážky však snižují efektivitu práce se systémem a využití zabezpečovacích či právě inovativních prvků uživatelského rozhraní. Mezi tyto inovativní prvky patří například

- oznamování chybně vyplněných polí formulářů hned po změně hodnoty,
- výrazná nápověda po najetí kurzorem myši, tzv. tooltip, s možností HTML formátování,
- zobrazení náhledu obsahu stránek v prvku stránkování seznamů,
- či možnost zamknutí systému, která je popsána v následující kapitole.

Systém byl z pohledu jednotného vzhledu a bezproblémové interpretace JavaScriptu testován v prohlížečích pro operační systém MS Windows, není-li uvedeno jinak: MS Internet Explorer 6.0, 7.0 a 8.0, v Mozilla Firefox 2.0.0.12 a 3.0.10 (i GNU/Linux), Opera 9.64 a 10 Alpha (i GNU/Linux) a konečně Google Chrome 1.0.154.65.

3.2.5 Bezpečnost

Data sledovaná a ukládaná systémem mohou být s ohledem na jejich citlivost například při organizaci reklamních kampaní pro konkurenci velice cenná. Systém tuto skutečnost zohledňuje v několika fázích.

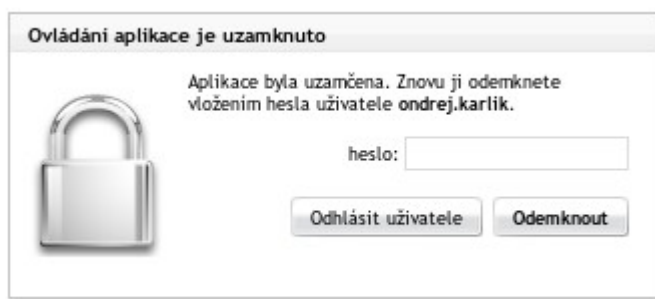
Každý neplatný pokus o přihlášení je monitorován, pokud jich uživatel v jednom dni zadá více jak 15, bude přístup z jeho IP adresy na dalších 24 hodin zablokován. Tímto je výrazně sníženo riziko útoku silou - opakováním kombinací znaků tak dlouho, dokud se útočník do systému nepřihlásí pod uhodnutými přihlašovacími údaji.

Má-li útočník možnost odposlouchávat data putující po síti uživatele, nešifrovaný přenos HTTP protokolu v případě prohlížeče se zapnutou podporou JavaScriptu obsahuje heslo v podobě *SHA-1 hashe* (heslo je posíláno konkatencí s veřejně přístupným řetězcem pro snížení rizika nalezení v databázi hashů), nedojde tak k prozrazení celého hesla, které by mohl poškozený uživatel používat na jiných serverech, zde však získaná data útočníkovi pro přístup do systému postačí.

Možností, jak celý přenos zabezpečit, je použití nadstavby HTTPS se šifrováním *SSL/TLS* - všechny odkazy v rámci systému jsou uváděny relativně, případné nasazení je tak velice snadné.

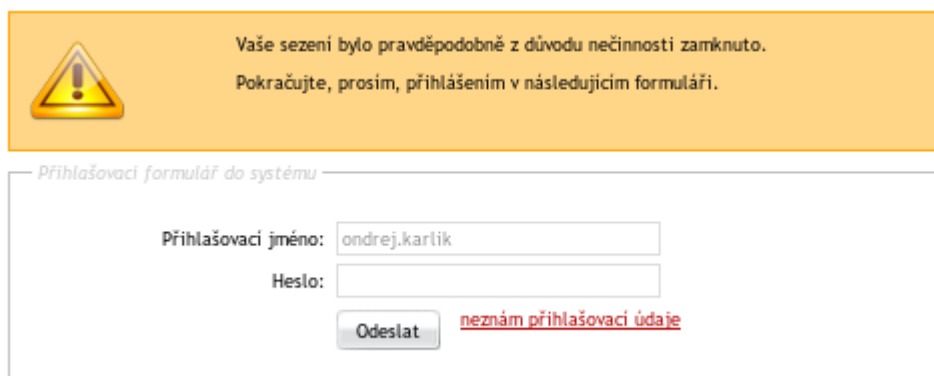
Riziko napadení systému již přihlášeným uživatelem je minimalizováno důsledným ošetřením vstupních dat - jednak všech dat vstupujících do SQL dotazů (útok zvaný *SQL injection*), jednak i řetězců zobrazovaných v HTML kódu (tzv. *cross-site scripting*). Útok známý pod zkratkou CSRF (v plném znění *Cross-site request forgery*) pak řeší vkládání speciálního řetězce do odkazů v rámci systému vázaného na dané sezení. Není-li pak v požadavcích daný řetězec přítomen, budou považovány za neplatné.

Riziko úniku dat za přímé fyzické účasti útočníka u uživatelova počítače snižuje automatické odhlašování ze systému. Toto odhlašování je však řešeno uživatelsky přívětivější formou - po šedesáti minutách nečinnosti (měřeno dle aktivity kurzoru myši či stisků klávesnice v okně webového prohlížeče) je uživateli zakryta obrazovka oknem se stručným vysvětlením a možností odblokování nečinnosti zadáním hesla - heslo je, znovu v podobě hashe, odesláno ke kontrole za pomoci technologie AJAX.



Obrázek 3.2: Formulář odemknutí aplikace AJAXem

Útočník může překryvnou vrstvu rozšiřovanou podle rozměrů okna například nástrojem *FireBug* odstranit či případně nahlédnout do zdrojového kódu, avšak na každé další stránce již uvidí obvyklý přihlašovací formulář. Dočasné blokování účtu bere v úvahu možnost otevření aplikace ve více prohlížečích zároveň (v tomto případě je blokováno každé sezení separátně) i ve více oknech či panelech jednoho prohlížeče, kdy se blokování řídí podle aktivity kurzoru či klávesnice v posledním použitém okně/panelu.



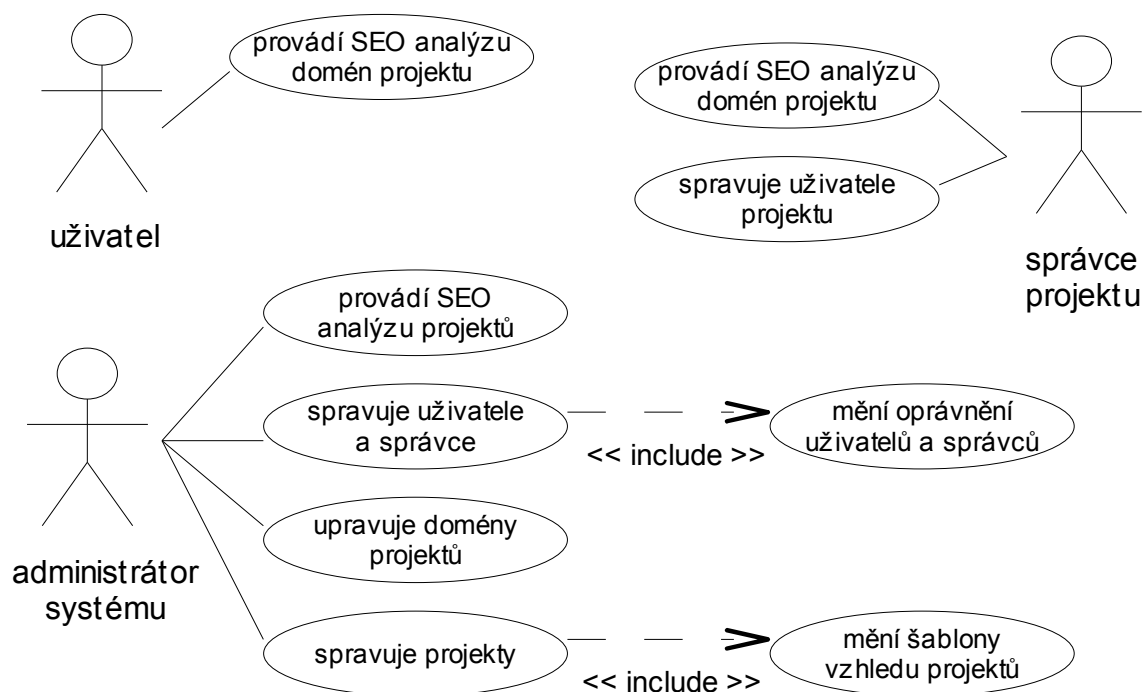
Obrázek 3.3: Obecný formulář odemknutí aplikace

3.2.6 Oprávnění

Systém byl navržen tak, aby umožňoval speciální hierarchii uživatelů a jejich oprávnění, totiž systém umožní nabídnutí služeb koncovým klientům, kteří si mohou SEO parametry svých stránek spravovat sami, a také společnostem, které se SEO analýzou zabývají. Provozovatel systému tak může pronajímat služby vázané na zvolené domény, aniž by uživatelé poznali, že používají produkt třetí strany - zde je tedy klíčovým požadavkem jednoduchá tvorba grafického uživatelského rozhraní, které bude možné jednotlivým klientům upravovat například do barev webu společnosti.

Administrátor systému, například tedy jeho vlastník, má neomezený přístup ke všem projektům i uživatelům, které může přidávat, mazat a nastavovat jim různé atributy. Projekty jsou myšleny domény či skupiny domén optimalizovaných stránek, uživatele v určených projektech pak mohou v rámci těchto domén využívat služeb systému - zjišťovat ranky, umístění ve vyhledávacích na zadaná klíčová slova apod. Uživatelé s rolí správce projektu pak mohou pro projekt vytvářet další uživatele, již však ne s rolí správce.

Systém oprávnění pak také umožňuje nastavení přístupu k jednotlivým službám, čehož může být využito například v případě zvýhodněného balíku pro firmy používající na některé služby jiné nástroje. Funguje na principu bitového pole podobného přidělování práv pro čtení, zápis a spouštění souborů v systémech Unix a od něj odvozených, avšak s rozdílem v užití řetězce, který svou délkou a tak i množstvím hodnot umožňuje více druhů oprávnění pro více akcí a služeb. Detailní popis funkce bitového pole čtenář najde v projektové dokumentaci.



Obrázek 3.4: use case diagram hlavních rolí v systému

3.3 Získávání dat

Nejpoužívanější vyhledávače zpravidla neposkytují přímo strukturovaná požadovaná data například ve formátu XML. Na řadu tedy přichází zpracovávání dostupného HTML kódu stránek s výsledky vyhledávání, případně využití externích nástrojů (např. validátorů zdrojového kódu). Systém splňuje požadavek modularity díky objektovému modelu, kde objekt je v tomto případě nazvaný *miner* (angl. horník, havíř). Tento název byl zvolen pro skutečnost, že se získávání dat potřebných pro systém dotýká problematiky *data mining*, tj. dolování dat z databází a jejich skládání do užitečných logických celků [10].

Každý objekt *miner* v systému implementuje metodu `mine()`, která vrací určenou strukturu dat v závislosti na jeho účelu. Tato struktura, obvykle v podobě pole, je pak zpracována a zobrazena v seznamech, grafech, sestavách či e-mailových zprávách s notifikací o změnách a výsledcích sledovaných parametrů v čase. Vzhledem ke skutečnosti, že systém umožňuje uživatelům jak manuální zadávání požadavků například na získání ranků sledované stránky, tak i automatické sledování jejich změn v čase, je nutné rozlišovat způsoby jejich zpracování - ručně zadané požadavky vyřizuje informační systém okamžitě, zatímco náročné požadavky, z nichž jsou vytvářeny například měsíční grafy vývoje ranků, jsou nejprve vloženy do fronty požadavků, odkud jsou poté ve zvoleném čase odebrány ke zpracování.

Do fronty se tyto požadavky ukládají ve formě záznamů s identifikací mineru, který je má zpracovat, a parametrů požadavku v tzv. *serializované* struktuře. Serializace umožní uložení pole do podoby řetězce a jeho následný převod zpět bez ohledu na počet zanořených polí, což uvádí výhodu univerzálního zpracování i z pohledu dlouhodobého vývoje systému - pole může jednoduchým přejmenováním klíče vytvořit novou větev s nekompatibilní strukturou, čemuž se *miner* v metodách zpracovávajících záznamy může snadno přizpůsobit.

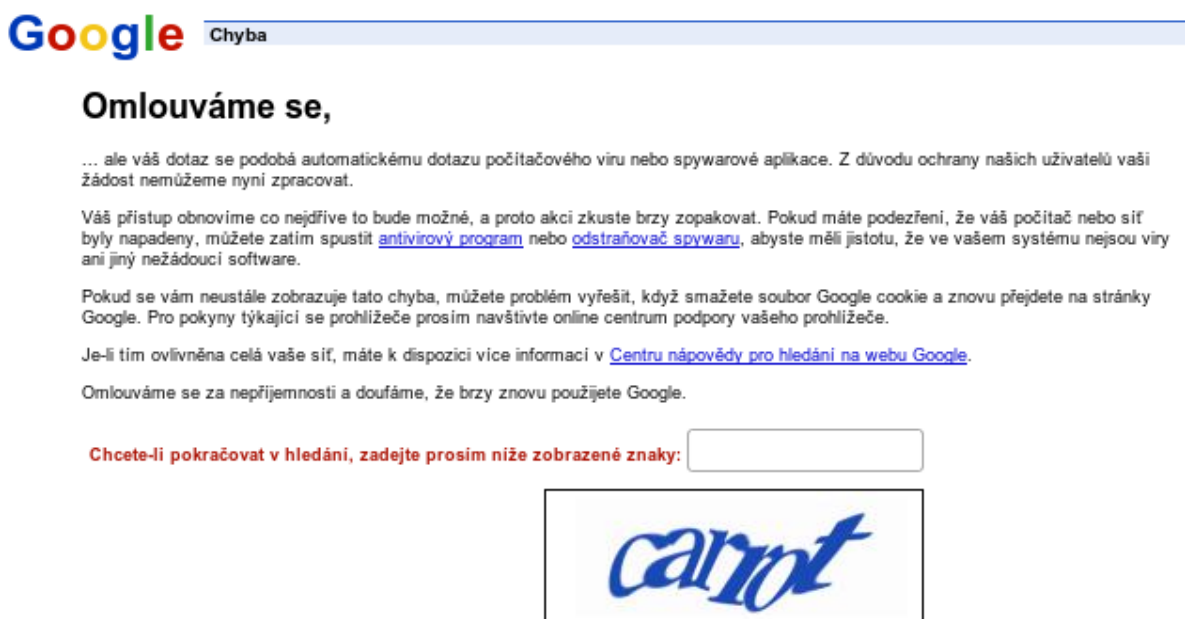
3.3.1 Problémy masového nasazení systému

Aplikace na jednorázové získání požadovaných dat spuštěná na počítači zákazníka může z jedné IP adresy při běžném použití generovat jednotky či desítky požadavků na vyhledávače za hodinu. Je-li však tento nástroj spuštěn na serveru a maximálně využíván, počet požadavků se může zvýšit až za vyhledávačem povolenou hranici.

Jednoduchý test, který jsem na výkonově průměrném osobním počítači s internetovým připojením o šířce pásma 10 mb/s provedl, ukázal, že opakované zpracování stránek (jejich stažení, kontrola přítomnosti řetězce a uložení výstupu) s výsledky vyhledávání serveru Google potřebuje přibližně 1 procento času procesoru a 200 kB/s (tj. 1,6 mb/s) pásma připojení. Nevyužití celého pásma může být způsobeno omezením pásma ze strany serveru vyhledávače na jedno spojení - při paralelním běhu testovacího skriptu pak využití lineárně narůstalo. Počítač zpracoval za uvedených podmínek průměrně cca 100 požadavků za minutu na každou spuštěnou instanci skriptu.

Účelem testu však bylo zjistit, kde se nachází zmíněná vyhledávačem povolená hranice počtu požadavků - jejich celkový počet se v mém případě vyšplhal k 9871 v rozsahu přibližně padesáti minut. Poté server Google automatizované činnosti zabránil. Zjištěný počet je bohužel jen orientační, protože druhou stranou měřený časový úsek mi není znám, oficiální informaci se mi najít nepodařilo.

Dalším aspektem, který je nutné brát v úvahu, je tedy výkon a vytížení serveru a také kvalita a šířka pásma připojení k Internetu.



Obrázek 3.5: Google - zamezení opakovaným požadavkům

3.3.2 Architektura a protokol

Vzhledem k nárůstu výkonu celého systému, snížení rizika zablokování automatizované činnosti i z důvodu vyšší spolehlivosti jsem se rozhodl problematiku řešit rozložením práce na více počítačů, klientů.

Distribuované zpracování fronty požadavků zde vyžaduje vytvoření protokolu komunikace mezi jednotlivými klienty, aby byly schopni domluvit se na tom, který z nich provede kterou část požadované práce, případně na zastoupení při výpadku jednoho z klientů či přerozdělení práce tehdy, když je jeden z počítačů přetížen například kvůli nestejně časově náročným požadavkům.

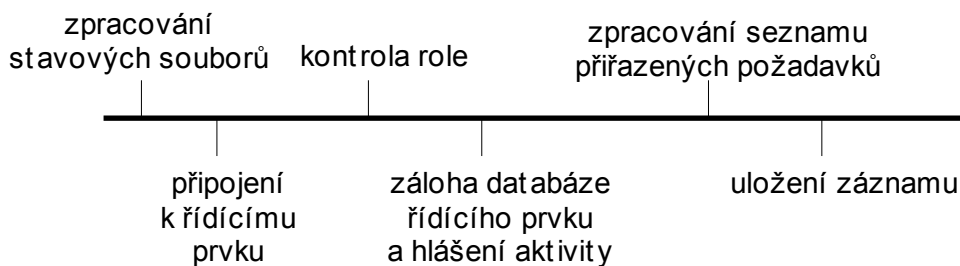
Uvažujeme-li soustavu klientů, tedy webových serverů, na nichž běží aplikace, ideálním způsobem, jak zajistit jejich komunikaci, je vytvoření prostředníka - řídicího prvku. Toto řešení s sebou však nese nepříjemný problém možnosti poruchy prostředníka a tím pádem zastavení činnosti všech připojených klientů.

Následující zjednodušený popis činnosti počítačů připojených do architektury vysvětluje i komunikaci mezi klienty, která neprobíhá přímým spojením, jak jej známe například z protokolu HTTP a jemu podobných, ale přes dostupné prostředky informačního systému, tedy vzdáleným připojením k databázi a zpracování skriptů jazyka PHP.

1. zodpovědná osoba, nejčastěji správce dostupných počítačů, již při instalaci zvolí, které počítače budou do řešení fronty požadavků zapojeny. Posléze těmto počítačům přidělí prioritu - ta je ztvárněna číselně a zároveň nutně unikátně, aby nedošlo k nejednoznačnostem při dalším využití. Vznikne tak seznam různě hodnocených serverů, kde kritériem je jejich spolehlivost a dostupnost, případně dostatečná výkonová rezerva pro případ převzetí role řídicího prvku.
2. Jak řídicí prvek do fronty požadavků ztvárněné tabulkou v jeho databázi vkládá nové akce s atributy frekvence provádění, identifikátorů minerů či například datem počátku provádění, ostatní klienti (zde s nižší prioritou) do fronty v pravidelných intervalech nahlíží a čekají na stav, kdy v ní naleznou požadavky právě jim přidělené. Při každém nahlédnutí do fronty zároveň řídicímu prvku oznamují, že jsou připraveni ke zpracovávání, aby řídicí prvek mohl tuto případně chybějící informaci zohlednit při přidělování požadavků - tato skutečnost tedy řeší výpadek jednoho z klientů, kdy řídicí prvek náhodně rozdělí jeho nedokončenou práci.
3. Úkolem všech klientů je také pravidelně zálohovat frontu požadavků i seznam ostatních počítačů do své lokální databáze. Při výpadku řídicího prvku tak může server s nižší prioritou roli řídicího prvku převzít, aniž by bylo nutné opakovat řešení velkého množství již jednou provedených požadavků - maximálně tedy těch, které byly provedeny od poslední zálohy, která ve výchozím nastavení probíhá v pětiminutových intervalech.

4. Klient výpadek řídicího prvku nejčastěji pozná nemožností připojení se k jeho databázi, kde problém identifikuje pomocí funkce jazyka PHP vracející jednoznačné číselné označení chyby připojení. V tomto případě se pak spojí s databází klienta s druhou nejnižší prioritou (či při analogických problémech třetí, čtvrtou atd.), kde tuto skutečnost oznámí. Pokud i tento klient zjistí problém s připojením, vyhláší sám sebe dočasným řídicím prvkem a připojení klienti začnou zpracovávat jeho frontu požadavků. Rizikem může být výpadek spojení mezi sítěmi, kdy problémy s připojením nemusí upozorovat všechny klientské počítače. Takovouto separaci skupin klientů aplikace v současnosti neumí řešit, avšak bezpečný návrat z této situace zajišťuje mechanismus kontroly databází podřízených serverů, v nichž může být v případě nalezení záznamu o degradaci řídicího prvku synchronizována databáze a nastolen stav s jediným (původním) řídicím prvkem. Obdobně se aplikace zachová po připojení upřednostňovaného klienta. Všechny chyby tohoto charakteru jsou posílány e-mailem na adresu správce serverů, který podle četnosti výpadků může rozhodnout o nucené degradaci některého z klientů či snížení jeho priority.
5. Všichni klienti, jak již bylo zmíněno v kapitole 3.1.1, znají rozdíl času oproti řídicímu prvku architektury a tento údaj promítají do ukládání záznamů o provedených požadavcích, aby se zamezilo nesrovnalostem při přepínání role řídicího prvku mezi dvěma počítači v jiných časových pásmech.

Následující časová osa ilustruje akce jednoho cyklu průběhu konečného automatu klientského počítače bez role řídícího prvku. Klient s touto rolí pak přidává operace přidělování a přerozdělování požadavků a další řídící operace. Problematika souborů pro změnu stavu je popsána v další kapitole.



Obrázek 3.6: cyklus aplikace pro zpracování dat

3.3.3 Výhody a nevýhody zvoleného řešení

Přestože některé klady a zápory již byly v předchozích kapitolách vysvětleny, shrňme je ještě jednou společně s dosud neuvedenými.

System rozdělením zpracování na více serverů s unikátními veřejnými IP adresami tolik neohroží možnost zablokování přístupu ze strany vyhledávačů a zároveň rozloží zátěž i spotřebu šířky

pásma připojení k Internetu, což může být výhodné pro společnosti, které mají ve správě více serverů specializovaných na různé projekty.

Řešení závislé na řídicím prvku může zkolabovat na jeho výpadku, čímž bez rychlého zastoupení novým řídicím prvkem ztrácí funkčnost celá síť klientů. V tomto případě je tedy maximální možný výpadek dlouhý pět minut - jde o kompromis mezi častým duplikováním obsahu databáze řídicího prvku a dlouhým výpadkem, kde by bylo nutné vykonávat znovu příliš mnoho akcí. Na druhou stranu návrh komunikace mezi klienty, která by byla plně decentralizovaná, by byl velice náročný - jak z hlediska rozsahu komunikace, která by musela pokrývat rozesílání zpráv o změnách ve frontě či v seznamu aktivních klientů, tak pro náročnost implementace rozhodovacích mechanismů pro přetřazování požadavků a způsobů rychlé aktualizace informací mezi všemi klienty včetně jejich připojování a odpojování.

Nevýhodou automatického spouštění jediného procesu zpracovávajícího požadavky je skutečnost, že v době běhu aplikace je soubor skriptu interpretem otevřen a tak nelze soubor aktualizovat přepsáním [11], takže by mohlo v určitých případech dojít k nekompatibilitě protokolu. Vhodným řešením je opakovaná kontrola existence speciálního souboru po každém zpracování požadavku. Přítomnost tohoto souboru v zadaném umístění pak způsobí ukončení vykonávání skriptu, čímž dojde k jeho aktualizaci. Aplikace dále využívá podobný soubor pro vynucení aktualizace seznamu připojených klientů.

Rozdělení zpracování mezi klientské počítače neřeší možnou nedostupnost webového serveru pro uživatele systému. Protože rozšíření informace o změně takzvaného DNS záznamu do celého světa může trvat i několik hodin, řešení ze strany informačního systému či řídicí aplikace není možné, prioritou je však zachování funkčnosti například generování sestav nebo pravidelných upozornění o změnách sledovaných parametrů.

Protože databázový a webový server nemusí být nutně instalován v jednom počítači, dá se jejich využití libovolně kombinovat - dokonce, za určitého nutného snížení spolehlivosti, lze spustit aplikaci na několika klientských počítačích, zatímco se všechny budou připojovat do databází fyzicky uložených na jednom serveru. Jde-li o vysoce optimalizovaný databázový stroj zapojený navíc do tzv. clusteru pro zvýšení spolehlivosti a výkonnosti, riziko výpadku dramaticky klesá a snižují se výkonové nároky na klientské počítače [12].

Mezi další kladné aspekty tohoto řešení patří i skutečnost, že je pro komunikaci využito stejných prostředků jako pro běh samotného rozhraní informačního systému, což je výhodné v případě nasazení pouze na jediného klienta, který bude zároveň řídicím prvkem.

3.4 Zpracování dat

OLAP, zkratka pro *On-Line Analytical Processing*, figuruje jako klíčový termín při zpracování získaných dat systémem. Konkrétně ve fázi sdělení informací analytikům - zpracované informace z takzvaného *datového úložiště* mají sloužit k podpoře rozhodování. Například při výběru klíčových slov podle vývoje a změn pozic ve vyhledávačích je úkolem OLAP a souvisejících technologií předat informace ne ve formě tisíců detailních záznamů, ale vhodně ve formě grafu či procentuálního vyjádření vývoje s rozdílem například oproti předchozímu sledovanému období, návštěvnosti stránek, případně jiným parametrům.

Procesu zhuštění těchto informací do přehledné formy se přímo týká sumarizace dat, která například zpracuje údaje z každého dne v týdnu a vrátí pak průměr hodnot za celý týden. Jde-li však o komplexní výpočty v rámci milionů záznamů, časová náročnost stoupá a musí být zavedeno řešení.

Detailní vysvětlení problematiky OLAP čtenář najde v uvedené literatuře [14], tato kapitola již uvede pouze popis řešených problémů s touto technologií souvisejících.

3.4.1 Sumarizace a agregace

Architektura objektů typu miner umožňuje těmto objektům jednoduše definovat možná kritéria pro sumarizaci (shrnutí nejen v podobě součtů) a agregaci (seskupení) - v případě implementovaného informačního systému jde o několik intervalů v čase, domén, skupin klíčových slov či skupin jiných sledovaných SEO parametrů - a určit jejich způsob i frekvenci.

A právě minery provádí většinu úkolů agregace a sumarizace, konkrétně tedy v podobě zpracování požadavků fronty, kdy mohou jednou za týden připravit data pro graf ročního vývoje ranků, vypočítat různé grafy pro přehled účinnosti optimalizace pro vyhledávače apod. Zbytek práce pak provedou funkce pro zobrazení grafů či na výpočet procentuálních hodnot změny.

Vzhledem ke způsobu ukládání výsledků atomických operací minerů do databáze, kdy mezi řetězci obsahujícími serializovaná pole nemůžeme jednoduše vytvořit vazby, vzniká pro většinu minerů v databázi vlastní tabulka, do nichž minery ukládají informace ve formě pro jednotlivé minery specifické, tedy například s uvedením data získání informace, úspěšností získání, frekvencí a kritériem již předem vypočítané agregace - počtem chyb zpracování v daném období či procentuálním vyjádřením změny vývoje hodnot ranků v určitém sledovaném intervalu. Seznamy či funkce generující grafy pro tyto minery pak již pracují s jednoduchými SQL dotazy, které jsou jednoduché na správu, zejména však vhodné na okamžité zobrazení jinak složitě získaných údajů.

Je vhodné dodat, že sumarizace a agregace může být využita již při ukládání detailních dat v původní podobě - uvažujeme-li například nové zpětné odkazy na doménu sledovaných stránek, je zbytečné ukládat jejich plný seznam či jejich počet.

3.4.2 Archivace dat

Každý miner při své práci v pravidelných intervalech maže velmi stará neagregovaná data v původním formátu serializovaného pole, protože se předpokládá, že analyticky v drtivé většině případů nebude zajímat výsledek konkrétního měření starého několik let, ale vystačí si například se zobrazením hodnoty v týdnu či měsíci nejbližšímu požadovanému datu.

U denně ukládaných záznamů tak může agregace záznamů do jednoho měsíce přinést uvolnění místa až třiceti záznamů (počet dní měsíce - 1 záznam zabraný sumarizací). Samotnou archivaci implementovaný systém neprovádí, měla by být, byť v omezeném rozsahu, součástí rutiny správy serverů.

3.4.3 Chybějící informace

Přestože jsou minery připraveny na situaci, kdy se například s některým ze zdrojů hodnot ranků nedokážou spojit z důvodu výpadku kontaktovaného serveru, bez interakce člověka není rozhodování o další činnosti úplným řešením - miner se v případě chyby pokusí akci provést znovu za 5 minut (a následně v dalších čtyřech pokusech v postupně trojnásobném časovém intervalu), avšak ani za několik hodin nemusí být druhá strana schopna odpovědět. Miner v tuto dobu pak také nemusí data již nutně potřebovat, tak snahu vzdá a ukládá chybu o nemožnosti získání požadované informace (a zároveň upozorní správce systému o chybě e-mailem).

Takováto chyba se musí promítnout v grafech a uživatel o ní musí vědět, aby na neplatné statistiky či hodnoty neproinvestoval peníze či úsilí. Seznam hodnot zobrazený v informačním systému tedy může uvést i počet chyb a dále zvýraznit nepřesnou hodnotu.

3.5 Umístění demonstrační verze systému

Informační systém byl pro náhled a představu o jeho schopnostech v aktuální verzi umístěn na adresu <http://seo.viaaurea.cz>. Zde poběží ještě nejméně rok, tedy do května 2010. V ostrém provozu, tedy se spuštěnou aplikací pro získávání a správu dat, bude aktivní pouze tento počítač.

Přihlašovací údaje do systému jsou z důvodu bezpečnosti umístěny na přiloženém CD v souboru *README.txt* v kořenovém adresáři disku.

4 Závěr

Výstupem mého snažení je informační systém postavený na jednoduchém univerzálním frameworku, který již plní poslání zjednodušení práce při analýze a optimalizaci pro vyhledávače například při řešení projektu v předmětu Tvorba webových stránek (ITW).

Díky velkému důrazu na modularitu systém umožňuje snadné rozšíření o sledování dalších parametrů. Mezi užitečná rozšíření může patřit například modul automatického návrhu vhodných klíčových slov na základě obsahu sledovaných i konkurenčních webových stránek, případně crawler, který bude analyzovat a hlásit duplicitu obsahu webových stránek, titulků stránek a meta-značek popisu a klíčových slov. Systém je zároveň připraven pro komerční použití, to díky navržené správě oprávnění i uvažovanému řešení problémů spjatých s masovým využitím.

Pro zvýšení výkonu celé aplikace plánuji vytvoření kompaktní verze frameworku bez komentářů a s celkovým zhuštěním zdrojového kódu v jediném souboru, čímž interpret jazyka PHP ušetří množství paměti i diskových operací. V budoucnu, konkrétně se zkušenostmi z předmětů navazujícího magisterského studia, bych se také rád soustředil na optimalizaci výkonu datového úložiště například rozložením zátěže nejen při získávání, ale i při zobrazení získaných informací, rozšířením protokolu komunikace mezi klientskými skripty, čímž by byl systém ještě výkonnější a spolehlivější.

Zkušeností s implementovaným frameworkem a samotnou tvorbou systému za pomoci nástroje Subversion pak využiji při tvorbě informačních systémů jiného druhu, protože dle mého názoru framework je, bez některých specializovaných aplikací třetích stran - například validátorů HTML a CSS -, dostatečně flexibilní, bezpečný a uživatelsky přívětivý.

Literatura

- [1] Baar, O.: *Robot = Crawler = Spider = Bot* [online].
Poslední modifikace 21. února 2009. [cit. 2009-05-01]. Dostupné na URL:
<<http://clanky.tez.cz/slovník-pojmu/robot-crawler-spider-bot.html>>
- [2] Thies, D.: *The Search Engine Marketing Kit* [online].
Poslední modifikace 6. dubna 2005. [cit. 2009-05-01]. Dostupné na URL:
<<http://www.sitepoint.com/article/search-engine-marketing-kit/>>
- [3] Navrcholu.cz: *Spojení Atlasu s Centrem pozici Seznamu ve vyhledávání neohrozilo* [online].
Tisková zpráva. Publikováno 15. července 2008. [cit. 2009-05-04]. Dostupné na URL:
<<http://www.iinfo.cz/tiskove-centrum/tiskove-zpravy/navrcholu-cerven-vyhledavace/>>
- [4] Neuman, P.: *Seznam: Jeden z 5 vyhledávačů, jež odolávají celosvětové expanzi Googlu* [online].
Poslední modifikace 21. září 2008. [cit. 2009-05-04]. Dostupné na URL:
<<http://buzzmag.cz/seznam-jeden-5-vyhledavacu-jez-odolavaji-googlu-21-09-2008/>>
- [5] Jyxo: *Služby* [online]. [cit. 2009-05-10]. Dostupné na URL: <<http://jyxo.cz/d/serv>>
- [6] Microsoft: *Seznam.cz použije pro vyhledávání v zahraničním internetu službu Live Search* [online].
Tisková zpráva. Publikováno 9. února 2009. [cit. 2009-05-04]. Dostupné na URL:
<http://www.microsoft.com/cze/presspass/MSG/20090209_news1.msp>
- [7] Bryant S.: *Hitwise: Google at 65 Percent of All Searches* [online].
Poslední modifikace 9. května 2007 [cit. 2009-05-04]. Dostupné na URL:
<http://googlewatch.eweek.com/content/google_search/hitwise_google_at_65_of_all_searches.html>
- [8] Fishkin R.: *Search engine ranking factors* [online].
Poslední modifikace 2. dubna 2007 [cit. 2009-05-05]. Dostupné na URL:
<<http://www.seomoz.org/article/search-ranking-factors>>
- [9] Smička R.: *Optimalizace pro vyhledávače - SEO* [online].
Poslední modifikace 9. června 2004 [cit. 2009-05-05]. Dostupné na URL:
<<http://seo.jasminka.cz/>>
- [10] Wikipedia: *Data mining* [online].
Poslední modifikace 19. února 2009 [cit. 2009-05-07]. Dostupné na URL:
<http://cs.wikipedia.org/wiki/Data_mining>
- [11] Vojnar T.: *Operační systémy - Správa souborů*, 2008, str. 23. [online].
[cit. 2009-05-10]. Dostupné na URL:
<<http://www.fit.vutbr.cz/study/courses/IOS/public/prednasky/ios-prednaska-04.pdf>>

- [12] MySQL: *MySQL 5.0 Reference Manual - MySQL Cluster* [online].
[cit. 2009-05-07]. Dostupné na URL:
<<http://dev.mysql.com/doc/refman/5.0/en/mysql-cluster.html>>
- [13] Bernard B.: *Úvod do architektury MVC* [online].
Poslední modifikace 7. května 2009 [cit. 2009-05-10]. Dostupné na URL:
<<http://zdrojak.root.cz/clanky/uvod-do-architektury-mvc/>>
- [14] Hruška T.: *Datová skladiště a technologie OLAP* [online].
Poslední modifikace 1. listopadu 2004 [cit. 2009-05-10]. Dostupné na URL:
<https://www.fit.vutbr.cz/~hruska/vyuka/INShistorie/pr07_soubory/frame.htm.cs>

Seznam příloh

Příloha 1. Návod k instalaci systému

Příloha 2. Vzor postupu při optimalizaci webové stránky

Příloha 3. CD se zdrojovými texty, programovou dokumentací a manuálem

Příloha 1: Návod k instalaci systému

Následující návod předpokládá nainstalovaný software uvedený v kapitole 3.2.1. Pokyny k instalaci uvedeného software závisí na cílové platformě, operačním systému a částečně i síťové topologii, kompletní popis není cílem tohoto textu. Návod popisuje instalaci jediné instance systému, propojení více klientů pro distribuované zpracování je detailně popsáno v manuálu.

1. Uživatel zkopíruje do připraveného podadresáře umístěného v adresáři, z něhož webový server čte vykonávané skripty a jiné dokumenty (tzv. *document root*), obsah adresáře *src* přiloženého CD.
2. Nastaví oprávnění plného přístupu do adresářů *tmp* pro webový server, respektive uživatele operačního systému, pod jehož účtem server běží.
3. Volitelně nastaví v konfiguraci webového serveru adresář *log* tak, aby do něj server zapisoval záznamy chyb či přístupů na stránky systému. Dále nastaví oprávnění jako v předchozím bodě.
4. Relativní cestu do vytvářeného adresáře v bodě 1 vzhledem k nastavené cestě k *document root* uvede v konfiguračním souboru *app/config/general.inc*. Pokud budou webové stránky informačního systému zobrazeny na vlastní doméně (tzv. *virtual host*), nastavovaná relativní cesta bude zpravidla reprezentována pouze znakem lomítka „/“.
5. V tomto konfiguračním souboru také nastaví přihlašovací údaje MySQL databáze.
6. Například za pomoci nástroje *phpMyAdmin* pak importuje skript *seo.sql* umístěný mezi soubory kopírovanými v prvním kroku do databáze, po úspěšném importu je vhodné skript vymazat.
7. Nastaví automatické spouštění skriptu *app/exec/cron.php* interpretem jazyka PHP na interval jedné minuty. Automatické spouštění skriptu bude nejčastěji, v případě operačního systému unixového typu, nastavováno v konfiguračním souboru */etc/crontab*. Dále nastaví oprávnění pro adresář *app/exec* jako v bodě 2 návodu.
8. Pokud ve webovém prohlížeči přistupujícím na adresu systému uživatel nenarazí na žádné chybové hlášení a bude-li přepínání se mezi stránkami systému zdánlivě fungovat, může přistoupit k instalaci prostředí pro běh aplikací v jazyce Java (*JRE*, *Java Runtime Environment*) potřebného pro spuštění CSS validátoru.
9. Instalace HTML validátoru pak probíhá nastavením podle vzorového konfiguračního souboru *framework/3rdParty/html-validator/httpd/conf/httpd.conf*. Zde je třeba upravit cesty k spustitelným souborům validátoru a tyto úpravy přenést do konfiguračního souboru webového serveru. Server je poté nutné restartovat.

Příloha 2: Vzor postupu při optimalizaci webových stránky

Pro snazší pochopení problematiky čtenářem uveďme jednoduchý smyšlený příklad optimalizace pro vyhledávače (SEO). Zadavatelem SEO je společnost historicky se orientující na trh kvalitních pracovních nástrojů, která chce - díky podobným výrobním postupům, s nimiž má zkušenosti - profitovat na vzrůstajícím zájmu zákazníků o profesionální sportovní potřeby. Situaci komplikuje návyk zákazníků kupovat si toto zboží v zahraničí (zejména v německy mluvících zemích), ten na druhou stranu značí, že na českém webu je pro další e-shop společnosti místo.

Hlavním bodem zájmu jsou pro společnost vlastní výrobky v oblasti horolezectví (karabiny, friendly, kotvy, kladky a blokanty), turistiky (trekové hole a přenosné sady nádobí) a zimních sportů (vázání, lavinové sondy a teleskopické hole). Ostatní sortiment sportovního vybavení uvedených oblastí společnost přeprořádá s nízkou výnosností na základě smluvních partnerství.

Předchozí odstavce předurčují několik zásadních rozhodnutí - firma v oboru sportovních potřeb nemůže spoléhat na povědomí o jejím názvu, vlastní název tedy prozatím nebude patřit mezi nejdůležitější z klíčových slov, naopak názvy přeprořávaných produktů mohou návštěvníky na webové stránky upozornit (a výhodné ceny na nich udržet). Vzhledem ke skutečnosti, že zadavatel již vlastní licenci internetového obchodu pro pracovní nástroje a tento e-shop je z hlediska SEO na vynikající úrovni, klesají náklady a zároveň je položen první základní kámen pro tvorbu kvalitních zpětných odkazů, tedy provázání stránek těchto dvou obchodů.

Dalším ze situací daných rozhodnutí je budoucí mezinárodní pojetí internetového obchodu. Stejně jako Češi aktuálně využívají nákupy v zahraničních obchodech s větším sortimentem, tržní podíl nákupů tohoto zboží obyvateli východnějších zemí právě v ČR podle analýzy trhu není nevýznamný. Zároveň pak bude využito schopnosti přepočtu do měny Euro, čímž se potencionální množství zákazníků násobí.

Jak již bylo uvedeno, prioritou pro společnost jsou vlastní výrobky - jejich obecné názvy včetně oblastí horolezectví a turistiky tvoří klíčová slova optimalizace. Sousedství „sportovní potřeby“ pak sice vyjadřuje obor nového podnikání, avšak vzhledem k zmíněnému zaměření není nejvhodnějším výrazem. Navíc je toto sousloví ve vyhledávacích vysoce konkurenční, což se projevuje existencí mnoha již optimalizovaných webových stránek, například i e-shopem na adrese www.sportovni-potreby.cz, a v tomto pojetí velice širokým spektrem nabízených potřeb k pro společnost nezajímavým sportům (in-line, basketbal, atletika apod.).

Zajímavým, moderním a dobře zapamatovatelným pojmem je *outdoor* slučující aktivní sporty často vyžadující profesionální vybavení. Zároveň jde o mezinárodní výraz, tedy vhodný pro volbu

domény webových stránek - nejvýhodnější doménou by jistě byla *outdoor.cz* společně s nejčastější zahraniční doménou nejvyššího řádu (TLD), tedy *outdoor.com*, ta je však samozřejmě pro velkou konkurenci obsazena. Ani spojení *outdoor-vybaveni* či *vybaveni-outdoor* není z tohoto důvodu uvažovatelné, avšak *outdoor-potreby* je u obou TLD a různých pořadí slov volné. Toto sousloví je navíc dostatečně univerzální a málo konkurenční.

The screenshot displays search results for the keyword "outdoor". It is divided into two main columns. The left column features a yellow banner with sponsored links for "Outdoor Shop", "Husky trička zdarma", and "Outdoor naplno". Below this, there are several organic search results, including "Outdoor - spolehlivý obchod pro outdoor", "Outdoor vybavení", "OUTDOOR.CZ", "Outdoor Expert", and "Úvodní stránka - Svět outdooru". The right column also has a yellow banner with sponsored links for "Firemní akce - Adrenalin", "Teambuilding", "BartSport.cz - Outdoor", "Deproc CZ", "outdoor shop", "Firemní akce-Outdoor", "Outdoor kurzy a eventy", and "Campingaz-Coleman.cz". At the bottom of the right column, there is a link for "Další sponzorované odkazy".

Obrázek: Výsledky hledání výrazu „outdoor“

Volená klíčová slova budou využita ve spuštěné reklamní kampani PPC systémů Sklik a Google AdWords, tím pádem se zde znovu vyplatí volit slova konkrétnější a méně konkurenční. Společně s názvy produktů a oblastmi jejich využití, výrazy *outdoor*, *vybavení* a méně *sportovní potřeby* tedy vzniká seznam slov, které bude nyní nutné ve vhodných pořadích, kombinacích a jazykových mutacích uvést do titulků stránek, nadpisů a lokálních i zpětných odkazů. Velkou výhodou bývají systémy firem provádějících SEO, které v rámci spravovaných optimalizovaných stránek udržují tématické vazby pomocí kvalitních zpětných odkazů, díky čemuž je pak pro nové stránky start na internetovém trhu mnohem rychlejší a jistější.