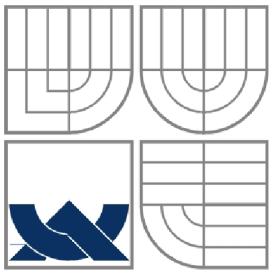
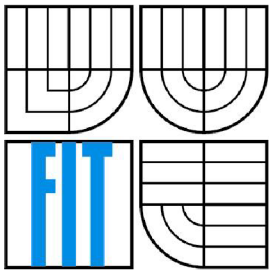


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

TECHNIKY INDEXOVÁNÍ WEBOVÝCH STRÁNEK

TECHNIQUES OF WEB PAGES INDEXING

BAKALÁŘSKÁ PRÁCE
BACHELOR´S THESIS

AUTOR PRÁCE — JIŘÍ TUŽIL
AUTHOR

VEDOUCÍ PRÁCE — ING. MICHAEL KUNC
SUPERVISOR

BRNO 2007

místo pro vložení Zadání bakalářské práce

Místo pro vložení kopie licenční smlouvy, s. 1

místo pro vložení kopie licenční smlouvy, s. 2

Abstrakt

Tato práce se zabývá způsoby vyhledávání informací v síti Internet. Popisuje strukturu prezentovaných dat a způsoby jejich převodu na informace, které je možné využít v procesu vyhledávání. Uvádí přístupy fulltextových algoritmů PageRank, HITS a SALSA, zmiňuje možná úskalí, nepřesnosti a zároveň výhody těchto vyhledávacích technik. Popisuje postup návrhu a implementace vzorového fulltextového vyhledávacího nástroje.

Klíčová slova

Internet, webové stránky, techniky indexování, techniky vyhledávání, vyhledávací algoritmy, vyhledávání informací, PageRank, HITS, SALSA

Abstract

This work addresses the techniques of information searching in the Internet. It describes the structure of the presented data and their conversion into information usable in searching process. It shows various approaches of PageRank, HITS and SALSA full-text algorithms, notifying about possible difficulties and inaccuracies as well as underlining the advantages of these search techniques. The work shows the design development and implementation of a sample full-text search tool.

Keywords

Internet, web pages, indexing techniques, techniques of searching, searching algorithms, information searching, PageRank, HITS, SALSA

Citace

Jiří Tužil: Techniky indexování webových stránek, bakalářská práce, Brno, FIT VUT v Brně, 2007

Techniky indexování webových stránek

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením Ing. Michaela Kunce a uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Jiří Tužil

Obsah

Seznam ilustrací

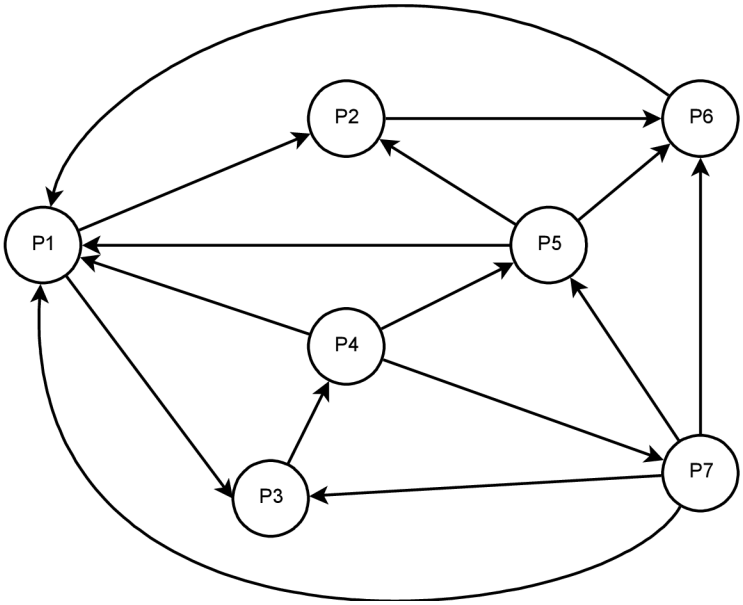
Seznam tabulek

1 Úvod

Internet se postupem doby a překotným vývojem stal díky svým takřka neomezeným publikačním schopnostem nejobsáhlejším informačním zdrojem dnešní doby. Množství informací, které zahrnuje, ale samo o sobě postrádá jakoukoliv ucelenou formalizovanou strukturu, podle které by bylo možné systematicky dohledat informace týkající se našeho zájmu. Neznáme-li přímou cestu k požadované informaci, nemáme žádnou jinou spolehlivou možnost, jak jí dosáhnout. I přes to, že mnohé zdroje informací v Internetu jsou mezi sebou vzájemně propojeny, nemůžeme předpokládat, že z každého známého zdroje existuje cesta k libovolnému jinému. A existuje-li, pak může být tak dlouhá, že úsilí vynaložené jejím následováním až k hledané informaci je v porovnání s přínosem nalezené informace neefektivní. Nemluvě o mnohých situacích, kdy až dokud to nenajdeme, přesně nevíme, co hledáme. Proto začaly v Internetu vznikat nové zdroje specializované na poskytování informací o tom, kde nalézt informace – vyhledávače.

1.1 Prostředí vyhledávačů

Velmi rozšířenou službou Internetu, která je využívána pro publikaci textových informací, je World Wide Web, zkráceně WWW nebo též jen web. Tímto označením jsou myšleny všechny soubory umístěné na HTTP serverech a vzájemně propojené odkazy.



Obrázek 1: Web jako orientovaný graf

Web si lze představit jako orientovaný graf jehož uzly představují soubory a hrany odkazy mezi nimi.

Soubory dostupné na WWW mohou mít jakýkoliv formát: text, obrázek, archiv, aplikace atd. My se budeme za účelem vyhledávání informací na webu zabývat blíže jen webovými stránkami – hypertextovými dokumenty ve formátu HTML.

HTML je tzv. značkovací jazyk, který párovými značkami (tagy) v toku čistého textu určuje jeho význam (sémantiku, např. kde začíná a končí nadpis nebo odstavec, kterou část textu chceme zdůraznit apod.) či nepárovými vkládá rozšířený obsah (obrázek, editační pole formuláře atd.).

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
"http://www.w3.org/TR/html4/strict.dtd">
<html>
<!-- toto je komentář -->
   <head>
      <title>Titulek stránky</title>
   </head>
<!-- tělo dokumentu -->
   <body>
   <h1>Nadpis stránky</h1>
   <p>Toto je tělo dokumentu, <a href="index.html">odkaz</a></p>
   <p><img src="nazev.souboru" alt="obrázek"></p>
   </body>
</html>
```

Vyhledávače webových stránek využívají při své práci maximum informací, které jim o souborech poskytuje struktura webu, přenosový protokol i (ale kupodivu ne především) jejich vlastní obsah.

1.2 Přehled typů vyhledávačů

Existuje několik druhů vyhledávačů, které se od sebe liší způsobem, jakým získávají a případně uchovávají informace o webových stránkách. Každý je vhodný na určité typy dotazů, není možné říci, že by jeden byl nejlepší.

1.2.1 Katalogové vyhledávače

Katalog nebo katalogový vyhledávač se podobá Zlatým stránkám – můžeme si ho představit jako kartotéku, do které jsou zařazovány webové stránky podle jejich oblasti zájmu. Kategorií, do kterých se zájmové oblasti dělí bývá velké množství a každá stránka obvykle obdrží ještě stručný popis. Není tedy problém najít rychle a přesně to, co hledáte – ovšem pouze tehdy, pokud se daná stránka v katalogu nachází a hledaná informace je dostatečně obecná na to, aby odpovídala kategorii nebo stručnému popisu.

Webové stránky jsou do katalogu zařazovány ručně jeho správcem. Protože je to časově náročná práce (jak bylo naznačeno v Úvodu), je třeba o zařazení vlastních stránek do katalogu zažádat a dodat správci jejich adresu URL, oblast zájmu a další požadované informace. Ten pak jenom ověří jejich správnost a Vaše stránka se ocitne v jedné nebo více kategoriích katalogového vyhledávače, dostupná chtivým hledačům.

Záznamy v katalogu jsou většinou nadále monitorovány a v případě dlouhodobé nedostupnosti vyřazeny, aby byl katalog udržován aktuální.

Katalogový vyhledávač je vhodný zejména pro vyhledávání firem, škol a dalších institucí, jejichž oblast působnosti lze přesně popsat a zařadit do kategorií katalogu. Nehodí se naopak pro hledání konkrétních slovních spojení nebo frází, které stručný popis v katalogu většinou nepostihuje.

- **výhody** přesnost výsledků
- **nevýhody** neúplnost, náročná manuální správa katalogu
- **používá** seznam.cz, centrum.cz, atlas.cz a mnoho dalších

1.2.2 Full-textové vyhledávače

Full-textový vyhledávač pracuje na zcela jiném principu. Zpravidla sestává ze tří oddělených částí – robota, který bez ustání prochází webové stránky odkaz za odkazem a sbírá jejich URL adresy, indexátoru, který nalezené stránky rozděluje na slova a ukládá jejich výskyty do

databáze, a vyhledávače, který je na dotaz uživatele schopný velmi rychle prohledat tuto databázi a zobrazit vhodné výsledky.

Full-textové vyhledávače se mohou chlubit vysokou úplností – jejich roboti a indexovací nástroje zaznamenají skutečně každou stránku, na kterou vede alespoň jeden odkaz z množiny již indexovaných stránek. To je ovšem vykoupeno vysokou strojovou náročností jak na výpočetní výkon a úložnou kapacitu, tak i velmi kvalitní konektivitu k Internetu, kterou indexování velkých objemů dat vyžaduje.

Full-textové vyhledávání je vhodné v případech, kdy hledáme přesný a nejlépe jednoznačný výraz. Ve výsledku vyhledávání obdržíme odkazy přímo na stránky s výskytem daných slov, nikoliv jen na vstupní stránku webu, jako je tomu u vyhledávání v katalogu.

- **výhody** úplnost i přesnost výsledků při vhodně zadaném dotazu
- **nevýhody** nepřehlednost výsledků, náročná implementace vyhledávacího stroje
- **používá** google.com, jyxo.cz, morfeo.cz, seznam.cz a mnoho dalších

1.2.3 Metavyhledávače

Metavyhledávač není vyhledávačem sám o sobě. Využívá výsledky vyhledávání z jiných vyhledávačů, které sloučí, seřadí podle počtu duplicit a zobrazí výsledek. Zároveň si udržuje žebříček úspěšnosti pro každý používaný vyhledávač a podle něj distribuuje příští dotazy, což napomáhá ke zvýšení relevance i rychlosti vyhledávání.

- **výhody** úplnost i přesnost výsledků
- **nevýhody** závislost na jiných vyhledávačích, pomalost
- **používá** vivisimo.com, metacrawler.com; z českých žádný

1.2.4 Rozšíření vyhledávačů

Vyhledávače jsou internetovými projekty mnohdy náročnými na provoz i údržbu a potřebují dosahovat nejen dlouhodobě kvalitních výsledků vyhledávání, ale také nabízet nové nadstandardní možnosti, aby si udržely náskok před konkurencí a tím i náklonnost zadavatelů reklamy a zdrojů financí.

Některé z rozšiřujících možností, které vyhledávače nabízejí:

- **seskupování výsledků podle tématu**
  snaha o kategorizaci výsledků full-textového vyhledávání, která zvyšuje jejich přehlednost na úroveň katalogu

- **grafické znázornění výsledků**
  prezentace výsledku jako množin či orientovaného grafu
- **náhledy webových stránek**
  spolu s textovými daty si vyhledávač ukládá i miniaturu grafické podoby každé stránky a zobrazí ji vedle výsledku vyhledávání
- **vyhledávání obrázků**
  vyhledávač při indexování uchovává i odkazy na obrázky a umí je vyhledat podle velikosti, rozměru nebo slov, která obrázek popisují
- **vyhledávání zboží**
  schopnost rozpoznat, vyhledat a porovnat položky z nabídky internetových obchodů, které používají známou obsahovou šablonu
- **vyhledávání článků a novinek**
  vyhledávač indexuje tzv. RSS feedy z internetových periodik a umí v nich samostatně vyhledávat podle titulku, popisu nebo času
- **vyhledávání v diskusních skupinách**
  vyhledávač se umí přihlásit do diskusní skupiny a prohledávat příspěvky
- **vyhledávání ve zdrojových kódech**
  vyhledávač „rozumí“ syntaxi programovacího jazyka a je schopen vyhledávat proměnné, funkce, třídy atd.

1.3 Srovnání vlastností předních českých vyhledávačů

Tabulka uvádí porovnání čtyřech českých vyhledávačů, které používají vlastní vyhledávací nástroje, z hlediska výše uvedených nadstandardních vlastností (stav k 5/2007):

	centrum.cz	jyxo.cz	seznam.cz	zacatek.cz
katalog	ano	ne	ano	ano
full-text	ano	ano	ano	ano*
meta	ne	ne	ne	ne
seskupování	ne	ano	ne	ne
graficky	ne	ne	ne	ne
náhledy	ne	ne	ano	ne
obrázky	ano	ano	ano	ano
zboží	ano	ano	ano	ano
zprávy / RSS	ano	ano	ano	ne

Tabulka 1: Porovnání nadstandardních vlastností českých vyhledávačů

Přestože Centrum.cz používá full-textový vyhledávač Morfeo.cz, jedná se o jeho dceřinou službu a budiž mu uznána za vlastní. Stejnou výjimku dostal i Seznam.cz u vyhledávání ve zprávách, jehož Novinky.cz po pár kliknutích vyhledávání nabízejí také. Zacatek.cz umožňuje full-textové vyhledávání pouze na stránkách uvedených ve svém katalogu a jejich podstránkách, ale i přesto je velmi dobře použitelné.

Jak je vidět, žádný z českých vyhledávačů nepodporuje alternativní grafické zobrazení výsledků, ani podporu pro své výsledky vyhledávání z jiného vyhledávače (což je ale z obchodního hlediska pochopitelné).

2 Teorie algoritmů

Protože algoritmy katalogového vyhledávače se omezují vesměs pouze na seřazení stránek uvnitř jedné kategorie (a to není nic složitějšího než spočítání výskytů hledaného výrazu v jejím popisu či častěji použití externího hodnocení stránky z jiného full-textového vyhledávače), budeme se nadále věnovat jen těm algoritmům hodnocení stránek, které v praxi uplatňují full-textové vyhledávače. Přestože jejich úplná implementace není známá a jedná se vždy o střežené duševní vlastnictví každé společnosti provozující vyhledávač, jsou díky výzkumným pracím autorů známy původní rysy nejpopulárnějších algoritmů. Nelze s určitostí říci, do jaké míry odpovídají těm skutečně používaným, je ale jisté, že některé postupy jsou obecně platné a používané, protože jejich vstupní veličiny vždy vycházejí z omezeného množství informací, které lze z webu o stránce pro danou část algoritmu získat.

2.1 PageRank

PageRank je iterativní algoritmus pro výpočet relativní hodnoty důležitosti dokumentu na základě počtu jeho referencí. Byl navržen v roce 1995 Larrym Pagem (od toho název Page-Rank) a Sergeyem Brinnem na Standfordské univerzitě jako výzkumný projekt a jeho funkční prototyp se stal základem hodnocení důležitosti webových stránek vyhledávačem Google. [4, 5, 6]

PageRank je založen na distribuci hodnocení samotné stránky dalším skrze odkazy. V první iteraci algoritmu má každá stránka libovolný počáteční PageRank, jehož poměrná část se v každé další iteraci přičte všem stránkám, na které odkazuje. Algoritmus je stabilní, rychle konverguje a výsledný PageRank má vždy hodnotu od nuly do jedné. Matematicky lze algoritmus zapsat jako

$$PR(a)=\frac{1-d}{N}+d\cdot\left(\sum_{u\in B_a}\frac{PR(u)}{N_u}+E(a)\right) \quad ,$$

kde $PR(a)$ je PageRank stránky a, d je tzv. dampening faktor (tlumící faktor), N je celkový počet uvažovaných stránek, B_a je množina všech stránek, které odkazují na a, N_u je počet odkazů, které vedou z u a $E(a)$ je počáteční zdroj PageRanku pro stránku a.

Součet PageRanku všech uvažovaných stránek je roven 1, lze tedy říci, že hodnota PageRanku představuje pravděpodobnost, se kterou se kliknutím na libovolný odkaz na libovolné z uvažovaných stránek dostaneme na stránku s tímto PageRankem (např. je-li $PR(A) = 0{,}5$, pak

je 50% pravděpodobnost, že se kliknutím na libovolný odkaz na dostaneme na stránku *A*). Dampening faktor *d* zařazuje do vzorce pravděpodobnost toho, že na nějaký odkaz vůbec klikneme, tj. že naše postupná cesta klikání po odkazech právě neskončí. Je stanoven experimentálně a obvyklá hodnota je asi 0,85.

Relativní výhodou PageRanku je možnost upřednostnění stránek oblíbeného tématu poskytnutím vyššího počátečního PageRanku a tím i jeho distribuci podobným relevantním stránkám. V praxi se tento přístup bohužel nepoužívá, protože výpočet personalizovaného PageRanku pro celou množinu stránek webu by byl výpočetně neúnosně náročný.

PageRank je možné počítat i pro ještě neznámé stránky a teprve podle jeho hodnoty rozhodnout, jestli má smysl je vůbec indexovat, popř. jak často.

Nevýhodou takto prezentovaného algoritmu je možnost favorizovat stránku velkým množstvím odkazů ze stránek s tématem, které odkazovaná stránka vůbec neobsahuje. Tohoto principu zneužívá GoogleBomb, technika, jak do výsledků vyhledávání zařadit zcela nesouvisející stránku jejím masovým odkazováním.

Algoritmus PageRank také upřednostňuje starší stránky na úkor mladších a to po dobu několika iterací, než hodnota PageRanku nově přidaných stránek začne konvergovat.

PageRank trpí i dalšími méně podstatnými nedostatky, jak ale dokazuje jeho praktická implementace ve vyhledávači Google, je možné je odstranit, potlačit, nebo nebo pominout, jsou-li dostatečně nepodstatné.

2.2 HITS

HITS (Hypertext Induced Topic Selection) je iterativní algoritmus navržený na Cornellské univerzitě Jonem Kleinbergem. HITS určuje pro každou stránku dvě hodnocení – *hub value* a *authority value*. [7, 8]

Hlouběji si všímá struktury webu, uvažuje, že stránky často působí jako rozcestníky (*hubs*), které odkazy soustřeďují, nebo stránky, které o tématu pojednávají, a na které je odkazováno (*authorities*). Tato hodnocení se vzájemně ovlivňují – *hub value* je součtem *authority value* stránek, na které odkazuje, a *authority value* je součtem *hub value* stránek, které na ni odkazují. Matematicky zapsáno

$$a_i^{(k)} = \sum_{j:e_{ji}\in E} h_j^{(k-1)} \quad , \quad h_i^{(k)} = \sum_{j:e_{ij}\in E} a_j^{(k)} \quad \text{pro k = 1, 2, 3, ...,}$$

kde a_i je *authority value* stránky i, h_j je *hub value* stránky j, e_{xy} je odkaz ze stránky x na stránku y, E je množina všech odkazů a k je pořadí iterace algoritmu. Je zřejmé, že v první iteraci musí být *hub value* i *authority value* každé stránky nastaveny na nenulovou počáteční hodnotu a tou je $1/n$, kde n je celkový počet stránek.

Od výše uvedeného algoritmu PageRank se HITS liší především v tom, že jeho výpočet je spouštěn při každém vyhledávacím dotazu. Přestože se nepočítá pro celou množinu stránek jako PageRank, ale jen pro relevantní dokumenty (vybrané jinou metodou), trvá jeho výpočet příliš dlouho na to, aby mohl být v takto prezentované formě prakticky použit. Upravený algoritmus HITS jako součást projektu CLEVER používá vyhledávač Ask.com.

2.3 SALSA

SALSA (Stochastic Approach for Link Structure Analysis) je iterativní algoritmus vyvinutý Ronny Lempelem a Shlomo Moranen na Israelském institutu technologie Technion. [8, 9]

Zatímco PageRank hodnotí všechny stránky jedním absolutním měřítkem a z nich poté vybírá relevantní dokumenty, HITS nejprve vybere množinu relevantních dokumentů a až poté jim přiřazuje dvojí hodnocení, SALSA kombinuje to dobré z obou těchto algoritmů – hodnotí všechny stránky dvojím měřítkem, ale až poté z nich vybírá relevantní dokumenty a to s vyšší úspěšností než PageRank a nižšími výpočetními nároky než HITS.

2.4 SEO

SEO je zkratka z anglického Search Engine Optimization, česky optimalizace pro vyhledávače. Jde o techniku, kterou se majitelé webových stránek snaží naplňovat kritéria hodnocení kvality vyhledávacími algoritmy a mít tak své stránky řazeny ve výsledcích jejich vyhledávání relevantních, ale příliš obecných výrazů, na co možná nejvyšších pozicích. Tato praxe je vyžadována především u komerčních internetových prezentací, u nichž vyšší pozice ve vyhledávači znamená vyšší návštěvnost stránky a tím i více potenciálních zákazníků.

Jak už bylo zmíněno, přesná funkce algoritmů a jejich role při určování konečné pozice ve výsledku vyhledávání je bedlivě střeženým tajemstvím provozovatele každého vyhledávače (právě proto, aby nemohlo pomocí SEO dojít ke znehodnocování jejich práce) jsou známy obecné faktory, ze kterých algoritmy vychází a které pozici stránky ve výsledku pozitivně či negativně ovlivňují [10]:

- **Vhodná, krátká a neměnná URL adresa**

Rozhoduje-li vyhledávač o pozici dvou obsahově zcela shodných stránek, přičemž URL adresa první z nich bude http://www.autobazar.cz/ford-focus-cerveny-2d-1997 a druhé http://www.autobazar.cz/nabidka.jsp?car=141, pak vyšší pozice při hledání výrazu „ford focus“ dosáhne stránka první, protože hledaná slova obsahuje navíc ve své URL adrese.

- **Budování zpětných odkazů na stránku**
  Významnost stránky stanovuje vyhledávač obecně podle počtu odkazů, které na ni vedou z jiných stránek. Čím je jich více, tím je stránka lépe ohodnocena. Odkazy na Vaši stránku je nejvhodnější umístit v katalozích, v diskusích na dané téma(!), tiskových zprávách apod.
- **Používání správného titulku (elementu *title*)**
  Titulek je prvním kontaktem návštěvníka s obsahem Vaší stránky. Měl by být stručný, přesný a výstižný – určitě ne „Úvodní stránka“ atp. Je zobrazován jako nadpis stránky ve výsledcích vyhledávání.
- **Používání meta tagu *description, keywords***
  Meta tagy desctiption a keywords dovolují blíže popsat obsah a klíčová slova stránky. Nezneužívejte jich, pokud se klíčová slova nebudou vyskytovat i v textu, bude stránka místo vyšších pozic naopak penalizována.
- **Správné používání doporučených značek (tagů)**
  Relevance hledaných výrazů stoupá s jejich důležitostí na stránce. To znamená, že slovo uvedené na stránce v nadpisu (elementu h1-6) získává pro stránku vyšší pozici než stejné slovo uvedené pouze v textu mimo nadpis.
- **Kvalitní a aktuální obsah**
  I přesto, že stránka nesplní všechny výše uvedené SEO techniky, může být řazena velmi vysoko, bude-li mít kvalitní, unikátní a aktuální obsah, získá si respekt ostatních a důležité zpětné odkazy postupem času sama. A o to jde především.

3 Návrh full-textového vyhledávacího nástroje

Při návrhu full-textového vyhledávacího nástroje si nejprve stanovíme požadavky na vstupní a výstupní veličiny, které má výsledný produkt splňovat a způsoby, jakými toho lze dosáhnout. Vybereme nejvýhodnější z nich a dekomponujeme ho na maximální počet vzájemně nekonkurujících si paralelních procesů. Ujasníme si vstupní a výstupní veličiny a jejich možné transformace v každém procesu. Ověříme, že výstupní veličiny jednoho procesu jsou zároveň vstupními veličinami procesu následujícího, s výjimkou výstupních veličin konečného procesu (produktu), které musí odpovídat prvotně stanovenému cíli.

3.1 Specifikace

Pro názornost se budeme při návrhu držet obecně používaného modelu full-textového vyhledávače s neiterativním algoritmem pro hodnocení relevance, což výrazně sníží nároky na výkon počítače. Z pohledu uživatele se jedná o aplikaci, které na vstupu zadá hledaný výraz a od níž obdrží množinu URL adres stránek (obecně objektů) uspořádanou podle klíče, který vyjadřuje míru vhodnosti (relevance) objektu k tomuto výrazu.

Aby aplikace mohla naplnit podmnožinu vyhovujících objektů, musí mít k dipozici předem připravenou úplnou množinu objektů, ze kterých bude vyhovující vybírat aplikací funkce, která každý z těchto objektů ohodnotí a přidělí mu klíč – míru vyhovění objektu hledanému výrazu.

Prvním úkolem je získat úplnou množinu URL adres objektů k prohledávání a druhým sestavit hodnoticí funkci.

3.2 Získání úplné množiny URL adres objektů

K získání úplné množiny URL adres můžeme zvolit několik přístupů:

- **zkoušení náhodných URL adres**
  velmi nízká úspěšnost správného tipu, velmi vysoká neúplnost
- **manuální vkládání URL adres jako do katalogů**
  náročné na údržbu, vysoká neúplnost
- **stahování seznamu registrovaných domén od správce a doplnění na URL adresu**
  ideálně získáme pouze vstupní stránku, vysoká neúplnost

- **získávání URL adres z odkazů na již známých stránkách**
  potřeba výchozí URL adresy, relativní náročnost implementace, vysoká úplnost

Ani jeden způsob není zcela vyhovující. První tři způsoby jsou náročné na údržbu a generují málo URL adres. Poslední je výhodný co do množství získatelných URL adres, sám ale potřebuje dodat alespoň jednu výchozí. Nabízí se tedy kombinace prvních tří možností, kterými získáme vstupní URL adresy, a z odkazů na nich potom čtvrtou metodou kaskádovitě další a další. I když se může zdát, že je toto spojení všech způsobů spolehlivé a získáme pomocí něho URL adresy všech objektů webu, není tomu tak. Jeho úplnost záleží na dodaných výchozích objektech – nepovedou-li z nich žádné odkazy mimo ně samé, žádné další URL adresy nezískáme. Taková situace je naštěstí dostatečně nepravděpodobná díky velké provázanosti objektů mezi sebou.

3.3 Sestavení hodnoticí funkce

Dále sestavíme funkci, která pro každý objekt dokáže stanovit jeho vhodnost (relevanci) k hledanému výrazu. Hledaný výraz uvažujeme jako uspořádanou *n*-tici slov, která jsou předmětem zájmu hledajícího, oddělených mezerou. Tento výraz rozdělíme na jednotlivá slova, dílčí hodnocení provedeme pro každé slovo zvlášť a jejich součet stanoví celkové hodnocení relevance objektu k hledanému výrazu.

Informací, které lze o objektu zjistit, ať už z jejího umístění ve struktuře webu, informací, které o něm sděluje HTTP server v hlavičce odpovědi, nebo z jeho obsahu samotného, je velké množství. Ty, které můžeme použít jako kritéria hodnocení relevance rozdělujeme do čtyř skupin [4]:

- **statické**
  informace nezávislé na hledaném slově
- **dynamické**
  informace závislé na hledaném slově
- **on-site**
  informace, které lze zjistit z obsahu objektu samotného
- **off-site**
  informace, které lze zjistit z obsahu objektů, na které odkazuje, nebo které odkazují na něj

Vybereme a zařadíme ta kritéria hodnocení relevance objektu $k_i(O, S)$, která jsou dle [2, 3] hodnocena jako nejvýznamnější, najdeme jejich obor hodnot a stanovíme jejich váhu v hodnocení objektu:

1. počet objektů cizích webů obsahujících hledané slovo v odkazu na tento objekt nebo poblíž něho
2. hledané slovo se nachází v URL adrese objektu
3. hledané slovo se nachází v titulku objektu, jedná-li se o dokument
4. hledané slovo se nachází v popisu nebo seznamu klíčových slov objektu, jedná-li se o dokument
5. počet výskytů hledaného slova v nadpisech v textu objektu, jedná-li se o dokument
6. počet výskytů hledaného slova v textovém obsahu objektu
7. počet kliknutí na URL adresu objektu uvedenou ve výsledku vyhledávání

Kritérium i	Min. hodnota M_i	Max. hodnota N_i	Váha w_i	Skupina
1	0	celkový počet indexovaných objektů	50,00%	dynamická, off-site
2	0	1	15,00%	dynamická, on-site
3	0	1	10,00%	dynamická, on-site
4	0	1	6,00%	dynamická, on-site
5	0	celkový počet slov v dokumentu	7,00%	dynamická, on-site
6	0	celkový počet slov v dokumentu	4,00%	dynamická, on-site
7	0	celkový počet zobrazení ve výsledcích	8,00%	statická, on-site

Tabulka 2: Parametry kritérií hodnocení objektu

Počet kritérií $m = 7$. Hodnota kritéria k_i pro slovo S a objekt O je definována slovně.

Vážený průměr hodnot kritérií k_i po normalizaci udává hodnotu relevance R_O objektu O pro každé jedno slovo S hledaného výrazu V a součet těchto relevancí R_O je pak roven hledané relevanci R objektu O k hledanému výrazu V:

$$V = \{S_1, S_2 \ldots S_n\} \qquad R_O(O, S) = \frac{\sum_{i=1}^{m} \left(\frac{k_i(O, S)}{N_i} \cdot w_i \right)}{\sum_{i=1}^{m} w_i} \qquad R(O, V) = \sum_{i=1}^{n} R_O(O, S_i)$$

3.4 Proces indexace

Bylo by velmi zdlouhavé, kdyby vyhledávač musel při každém dotazu procházet a hodnotit všechny webové objekty znovu a znovu. Tuto činnost stačí provést pro každý objekt jednou, získat a lokálně uložit informace potřebné ke stanovení hodnot kritérií pro výpočet relevance a při vyhledávání je využívat. Bez tohoto kroku by byl full-textový vyhledávač kvůli své pomalosti prakticky nepoužitelný. Tímto jsme našli první nezávislý proces – proces indexace.

Indexací zde rozumíme získání a lokální uložení informací, potřebných k výpočtu hodnoticích kritérií, která jsme uvedli v předchozí kapitole. Tento proces tedy bude mít na starosti získání každého dosud neindexovaného nebo dlouho neindexovaného objektu z úplné množiny URL adres, jeho indexaci závislou na jeho typu a zároveň doplňování úplné množiny URL adres odkazy ze získaných objektů. Pro dokumenty v jazyce HTML zahrnuje indexace rozklad na značky (tagy) a slova textu, jejich ohodnocení a uložení do databáze slov, pro obrázky zjištění jejich formátu a rozměru a uložení do databáze objektů, pro emaily existenci domény apod.

Proces indexace se také stará o plánování reindexací. Ke každé URL adrese objektu přiřazuje datum příští indexace, které stanovuje podle „stáří“ objektu (data poslední změny).

- **Vstupní veličiny** množina URL adres objektů
- **Výstupní veličiny** množina slov s kritérii, množina URL adres nových objektů

3.5 Proces rekalkulace

Proces indexace přinesl informace pro výpočet kritérií. Stejně jako v předchozím případě, některé z nich není nutné vypočítávat znovu a znovu při každém vyhledávacím dotazu. Vytvoříme proto další samostatný proces – proces rekalkulace. Ten bude mít za úkol jednorázově vypočítat z informací získaných procesem indexace hodnoty statických kritérií pro každý možný pár stránky a slova a tyto hodnoty lokálně uložit.

- **Vstupní veličiny** množina slov s kritérii, množina URL adres objektů
- **Výstupní veličiny** množina relevancí pro každý pár slova a objektu

3.6 Proces hledání

Činnost, kterou jsme ještě žádným procesem nepostihli, je interakce s uživatelem. Pro ni vytvoříme třetí proces – proces vyhledávání. Ten převezme hledaný výraz od uživatele,

vypočte k němu hodnotu relevance každé stránky (výše uvedeným postupem s využitím předem vypočtených statických kritérií i okamžitě vypočtených dynamických kritérií) a vypíše URL adresy stránek seřazené podle tohoto klíče.

V jeho režii také proběhne inkrementace čítačů zobrazení URL adres objektů a kliknutí na ně.

- **Vstupní veličiny** uspořádaná n-tice slov hledaného výrazu, množina relevancí pro každý pár slova a objektu
- **Výstupní veličiny** množina URL adres objektů, uspořádaná podle výsledku hodnoticí funkce

3.7 Shrnutí

Navrhli jsme full-textový vyhledávač s neiterativním algoritmem hodnocení relevance objektů, stanovili hodnoticí funkci a jeho činnost rozdělili do tří vzájemně nekonkurujících si procesů. Ověřili jsme, že vstupy jednoho procesu jsou zároveň výstupy procesu předcházejícího s výjimkou posledního procesu, jehož výstupem je požadovaná množina URL adres objektů, uspořádaná podle klíče – výsledku hodnoticí funkce.

4 Implementace

Od návrhu k prvnímu prakticky použitelnému produktu vede ještě dlouhá cesta implementace. Na jejím začátku zvážíme možná implementační rozhraní a vybereme nejvhodnější platformu, programovací jazyk a databázový server. Sestavíme databázi podle předem specifikovaných požadavků, navrhneme strukturu pro sestavení zdrojového kódu a implementujeme tři nezávislé procesy. Každý z nich otestujeme samostatně i ve spojení s ostatními a uvedeme několik příkladů chování vyhledávacího nástroje nad vzorovými daty.

Pro lepší orientaci v postupu implementace je vhodné projít Uživatelskou příručku již hotové aplikace, která je publikována v *Příloze B* této práce.

Pro nižší náročnost na výpočetní výkon i úložnou kapacitu implementujeme pouze lokální full-textový vyhledávač, tj. takový, který indexuje pouze správcem zadanou stránku a všechny její podstránky.

4.1 Prostředí

S ohledem na cílení této webové aplikace byl výběr implementačního prostředí omezen následujícími kritérii:

- dostupnost na širokém množství platforem a operačních systémů
- rozšířenost a z toho plynoucí obecně dobrou znalost instalace a správy
- možnost instalace a provozu v omezeném prostředí komerčního webhostingu
- předpřipravenost pro práci v prostředí webu
- možnost získání alespoň nekomerčních licencí bezplatně

Výběr jsme zvažovali pro každou část aplikace samostatně. Pro webová rozhraní se jako jasná varianta jeví jazyk PHP, který je velmi rozšířený a podporovaný a v daném účelu vyhoví všem kritériím. Databázový server přicházel z úvahu MySQL nebo PostgreSQL. PostgreSQL vychází lépe v možnostech, MySQL vítězí na poli šíře podpory na komerčních hostinzích. S tabulkami InnoDB přichází i podpora referenční integrity a indexů nad více poli, včetně textových, které tabulkám MyISAM v nižších verzích MySQL chyběly, a to dostatečně přibližuje možnosti MySQL PostgreSQL. Abychom nebyli zcela závislí na použití jednoho konkrétního databázovém serveru, bylo na aplikační úrovni implementováno rozhraní, které umožňuje pracovat (přinejmenším) s oběma z uvedených.
Produkované webové stránky jsou validními dokumenty dle specifikace XHTML 1.1.
Vzorová implementace a testování probíhaly na databázovém serveru MySQL verze 4.1.9

a 5.0.37, PHP interpretru verze 4.4.0 a 5.2.0 s rozšířením `mbstrings`, HTTP serveru Apache verze 1.3.33 a 2.0.52, OS Linux Slackware a Microsoft Windows XP a prohlížečích Opera verze 8 a 9 a Mozilla Firefox verze 1.5.0.11, ve všech případech řádně a bez problémů.

4.2 Návrh databáze

V kapitole *Návrh full-textového vyhledávacího stroje* jsme našli veličiny, které budou jednotlivé procesy získávat a ukládat do databáze a jiné z ní čerpat. Z těchto veličin sestavíme tabulky, stanovíme referenční integritu mezi nimi, databázi zakreslíme do ER diagramu a vytvoříme SQL skript pro založení její struktury.

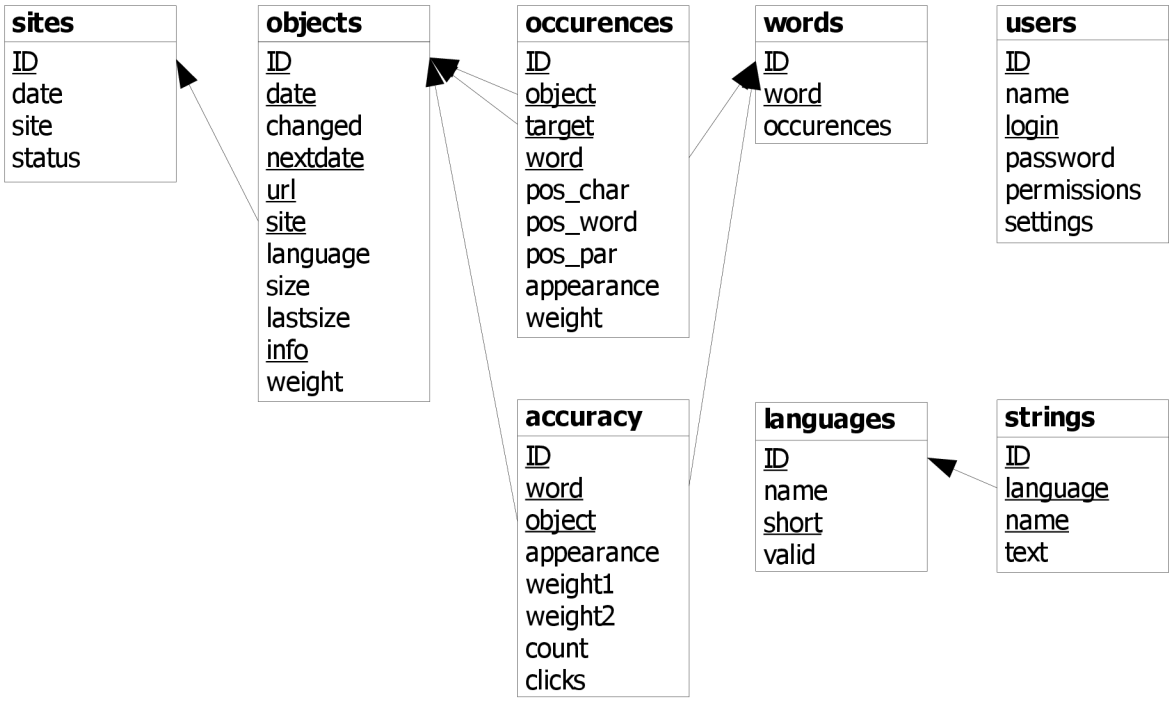
Toto jsou veličiny, které produkují procesy indexace a rekalkulace:

Množina URL adres objektů:	**Množina slov s kritérii:**	**Množina párů slovo/objekt:**
URL adresa	slovo	slovo
datum poslední změny	objekt	objekt
datum poslední indexace	obsaženo v URL	hodnocení
datum příští indexace	obsaženo v titulku	
velikost	obsaženo v popisu	
poslední známá velikost	obsaženo v nadpisech	
typ	obsaženo v textu	
jazyk	je odkazem na URL	
počet zobrazení ve výsledcích		

Tabulka 3: Veličiny produkované procesy indexace a rekalkulace

Uspořádání do databázových tabulek se od výše uvedeného výčtu příliš neodlišuje, pouze jsme rozdělili množinu slov s kritérii do tabulky unikátních slov a tabulky jejich výskytů, doplnili o další veličiny a přepracovali formát uložení dat. Bylo také třeba doplnit tabulku vstupních bodů webů a tabulky pro uchovávání uživatelských účtů, jazyků a textových řetězců, které podporují správu a multijazyčné prostředí vyhledávače.

Rozšíření tabulky výskytů slov (`occurences`) o pole `pos_char`, `pos_word` a `pos_par` je z důvodu budoucího rozšíření vyhledávače o možnost vyhledávání pomocí frází a operátorů bez nutnosti nové reindexace všech objektů v databázi. Aby bylo možné v hledaném výrazu specifikovat i význam slova podle elementů, ve kterých je obsaženo, převedli jsme jeho výskyt v kontrolovaných elementech na bitové pole `appearance` a zároveň přidali pole `weight`, které obsahuje součet vah těchto elementů. Pro zachování referenční integrity dat byly vytvořeny cizí klíče tak, jak naznačují šipky v diagramu. Nad podtrženými poli byly založeny indexy pro rychlejší spojování a vyhledávání v tabulkách.



Obrázek 2: ER diagram struktury databáze

Pro detailní specifikaci databáze viz *Přílohu A: Struktura databáze*. SQL skript pro vytvoření výchozí struktury databáze v dialektu MySQL naleznete na přiloženém CD, které je nedílnou součástí této práce.

4.3 Adresářová struktura a parametry zdrojového kódu

V projektu jsme implementovali tyto samostatné části:

- webové rozhraní pro správu vyhledávače (administrační rozhraní)
- skript pro indexování webu
- webové rozhraní pro vyhledávání a prezentaci výsledků (uživatelské rozhraní)

Každá z nich je umístěna v samostatném adresáři a má vlastní konfigurační soubor. Navíc jsme vytvořili adresář pro společné části kódu a společné konfigurační soubory.

Výsledná adresářová struktura vypadá takto:

```
/ -+- admin/          (adresář administračního rozhraní)
   +- config/         (adresář pro společné konfigurační soubory)
   +- include/        (adresář pro společné části kódu)
   +- spider/         (adresář se skripty pro indexování webu)
   +- web/            (adresář uživatelského rozhraní)
```

V kořenovém adresáři se nachází soubor `index.php`, který prohlížeč uživatele přesměruje k uživatelskému rozhraní do adresáře `web/`, pokud nespecifikoval jinak.

Protože aplikace nativně pracuje v kódování UTF-8, jsou i veškeré zdrojové soubory ukládány v tomto kódování.

4.4 Webová rozhraní

Při implementaci administračního a uživatelského rozhraní bylo zároveň vyvinuto unifikované prostředí pro malý informační systém, které se vyznačuje těmito částmi:

- přihlášení a odhlášení uživatele a detailní kontrola jeho oprávnění
- logické rozdělení spravovaných objektů na samostatné stránky a intuitivní navigace mezi nimi pomocí roletového menu, tlačítek a ikon
- propracované zobrazení a editace dat pomocí předpřipravených komponent tabulek, formulářů a dialogů

4.4.1 Sestavení stránky

Sestavení stránky unifikovaného webového rozhraní sestává z těchto kroků:

- přilinkování lokálního konfiguračního souboru `_config.inc.php`
- přilinkování společných částí zdrojového kódu `functions.inc.php` a `parts.inc.php`
- vytvoření vlastní PHP Session
- zabezpečení vstupních dat v polích `$_GET` a `$_POST` proti SQL-injection
- kontrola zalogovaného uživatele a případné přesměrování na logovací stránku
- přilinkování konfiguračního souboru databázové vrstvy

- přilinkování příslušného databázového rozhraní
- připojení k SQL serveru a databázi
- uložení uživatelského nastavení
- přilinkování hlavičky HTML souboru
- přilinkování roletového menu
- vlastní tělo každé stránky unifikovaného webového rozhraní
- přilinkování patičky HTML souboru
- odpojení od databáze a SQL serveru
- ukončení PHP Session

Tento princip výstavby stránek zaručuje přehlednost, nízkou duplicitu kódu, modulární koncepci reprezentovanou formátem přilinkovávaných souborů a přispívá k udržení validity produkovaných (X)HTML dokumentů tím, že minimalizuje množství zdrojového kódu nutné k napsání kvalitní webové stránky informačního systému. Díky předpřipraveným parametrizovatelným komponentám pro databázové tabulky, dialogy, tlačítka a formuláře je výstavba takovéto stránky dílem zavolání několika málo funkcí.

4.5 Webové rozhraní pro správu vyhledávače

Webové rozhraní pro správu vyhledávače je vystavěno nad unifikovaným webovým rozhraním. Umožňuje spravovat tyto entity:

- indexované weby a URL jejich vstupních bodů
- statistiky a výsledky indexování
- uživatelské účty a oprávnění
- jazyky a překlady textů
- zálohování

Pro detailní informace o webovém rozhraní pro správu vyhledávače viz *Přílohu B: Uživatelská příručka*.

4.6 Skript pro indexování webu

Skript pro indexování webu je navržen modulárně ze dvou hledisek – jednak podle podporovaných protokolů a poté podle podporovaných typů souboru. Moduly protokolů mají název `fetch_<protokol>.inc.php`. Jejich úkolem je otevřít předanou lokaci jako datový proud a vyšší vrstvě poskytovat na požádání data. Moduly typů souborů mají název `parse_<typ>.inc.php` a jejich úkolem je číst data souboru a získávat z nich informace potřebné pro jeho indexaci.

V základní implementaci jsou dostupné tyto moduly protokolů:

- `fetch_file.inc.php` – modul pro čtení souborů z lokálního souborového systému
- `fetch_ftp.inc.php` – modul pro čtení souborů z FTP serveru
- `fetch_http.inc.php` – modul pro čtení souborů z HTTP serveru
- `fetch_mail.inc.php` – modul pro čtení fiktivního protokolu mailto

A tyto moduly typů souborů:

- `parse_html.inc.php` – modul pro indexování dokumentů formátu (X)HTML / XML
- `parse_image.inc.php` – modul pro indexování obrázků formátu PNG / GIF / JPEG
- `parse_pdf.inc.php` – modul pro indexování dokumentů formátu PDF do verze 1.3

Přiřazení URL adres objektů k modulům protokolů probíhá na základě protokolu v URL uvedeném. Přiřazení souborů k modulům typů souborů probíhá na základě typu souboru, který vrátil modul protokolu. V praxi se jedná o MIME typ, který vrátí HTTP server v hlavičce, nebo je stanoven podle obsahu nebo přípony souboru.

Textové řetězce, které produkují moduly typů souborů před zápisem do databáze procházejí překódováním znakové stránky ze kódování objektu do kódování UTF-8. Tuto konverzi zajišťuje PHP knihovna `mbstrings`.

4.6.1 Úklid databáze

Úklid databáze spočívá pouze v kontrole vstupních bodů indexovaných webů – každý indexovaný web musí mít alespoň jeden objekt v tabulce objects. Nemá-li, je vytvořen s URL adresou vstupního bodu webu.

Dříve tento skript zastával mnohem více činností, které mohly být eliminovány se zavedením referenční integrity mezi tabulkami databáze a definicí pravidel `ON DELETE CASCADE` a `ON UPDATE CASCADE` pro tyto reference.

Skript se nachází v souboru `/spider/db_clean.inc.php` a je možné ho spustit jak samostatně, tak voláním include z jiného skriptu. Standardně je volán ze skriptu `/spider/spider.php` před spuštěním procesu indexace.

4.6.2 Indexování nových a zastaralých objektů

Hlavní částí skriptu pro indexování webu je vlastní indexace nových a zastaralých objektů.

Tento skript prochází databází objektů a zpracovává ty, které dosud nebyly indexovány, nebo ty, které mají prošlé datum příští indexace. Podle protokolu v URL adrese objektu je zavolán příslušný modul protokolu, který, pokud uspěje, poskytne datový stream obsahu objektu a údaje o objektu, jako jeho velikost, datum poslední změny, typ, jazyk atd.

Podle vráceného nebo stanoveného typu souboru je dále volán modul typu souboru, který indexuje jeho obsah a uloží do databáze. V případě modulu pro indexování HTML dokumentů se jedná o lexikální analýzu textu realizovanou konečným automatem, v případě indexování obrázků typu GIF / JPEG / PNG se jedná o zjištění jejich rozměrů a uložení jako rozšířeného typu objektu a v případě indexování souborů PDF se jedná o jeho dekódování a prosté rozdělení textu souboru na slova prázdnými znaky.

Každé slovo, které moduly načtou, předají indexační funkci k uložení. Ta se stará o přiložení bitového pole `appearance` stanoveného okolními elementy, výpočet váhy `weight` a aktualizování čítačů polohy slova `pos_byte`, `pos_word` a `pos_par`.

Narazí-li lexikální analyzátor modulu pro indexování HTML dokumentů na element, který obsahuje odkaz na libovolný jiný objekt, předá ho funkci pro zpracování odkazů. Ta převede relativní odkaz na absolutní, zkontroluje, zda-li patří právě indexovanému webu (shoduje se počátek URL adresy se vstupním bodem webu) a v případě, že by byl v databázi objektů unikátní, uloží ho. Při příštím průchodu skriptu pro indexování webu databází bude tento objekt indexován.

Při jednom spuštění prochází skript databázi tolikrát, dokud nachází alespoň jeden dosud neindexovaný nebo zastaralý objekt.

4.6.3 Rekalkulace databáze

Rekalkulace databáze odpovídá činnosti popsané v kapitole *Proces rekalkulace* – pomocí agregačních SQL dotazů nad tabulkami objects, words a occurences je naplněna tabulka accuracy, která je hlavním zdrojem informací pro proces vyhledávání. Tato akce je velmi výpočetně náročná a pro větší databáze v řádu desítek tisíců objektů může trvat i několik hodin(!).

Skript, který vykonává tuto činnost je `/spider/db_recalc.inc.php` a je možné ho spustit jak samostatně, tak voláním include z jiného skriptu. Standardně je volán ze skriptu `/spider/spider.php` po ukončení procesu indexace.

4.6.4 Automatické spouštění reindexace

Pro automatické spouštění reindexace slouží skript `/spider/reindex` (resp. `/spider/reindex.bat` pro OS Windows), který spustí vlastní reindexaci a přesměruje její výstup do databáze. Spuštění tohoto skriptu je vhodné nastavit do plánovače úloh v operačním systému a to v libovolných intervalech.

4.7 Webové rozhraní pro vyhledávání a prezentaci výsledků

Webové rozhraní pro vyhledávání a prezentaci výsledků je taktéž založeno na unifikovaném webovém rozhraní prezentovaném v kapitole *Webová rozhraní*. Prakticky se jedná o jedinou stránku, která obsahuje formulářová pole pro zadání vyhledávaného výrazu, resp. dalších vlastností vyhledávání jako jsou:

- hledání pouze ve vybraných elementech
- hledání jen objektů daného typu
- hledání jen těch objektů, jejichž URL obsahuje zadaný řetězec
- hledání objektů ve specifikovaných jazycích
- možnost skloňování českých slov
- možnost doplnění nebo odstranění diakritiky

Kapitola Proces hledání zřejmě naznačila postup, kterým vyhledávání probíhá:

- hledaný výraz je rozdělen na jednotlivá slova

- tato slova jsou pomocí služby pravopis.cz doplněna o všechny další dostupné české tvary, o (pa)tvary bez diakritiky a jedno slovo vytvořené konkatenací všech slov původního výrazu
- z takto upraveného výrazu je sestaven jediný SQL dotaz, který vrací množinu 10ti URL adres objektů uspořádanou podle jejich váhy k hledaným slovům; nalezené objekty musí obsahovat alespoň jedno slovo vyhledávaného výrazu, splňovat všechny nastavené rozšířené vlastnosti vyhledávání a zároveň i maximum kritérií uvedených v kapitole *Sestavení hodnoticí funkce*
- výsledek dotazu je vypsán do těla stránky jako seznam odkazů velikostí písma přímo úměrnou k hodnocení příslušného objektu
- každý odkaz ve výsledku vyhledávání vede na mezistránku redir.php, která zajistí inkrementaci čítače kliknutí na URL adresu objektu ve výsledku vyhledávání a na tuto adresu prohlížeč přesměruje

Lišta s rozšiřujícím nastavením vyhledávání je uživateli zobrazena až po té, kdy není spokojen s prvními výsledky vyhledávání. Velikost samotné vstupní stránky vyhledávače je díky tomu jen 1,7 kB.

Ačkoliv výpis výsledků vyhledávání může působit v některých situacích neprakticky, jedná se přinejmenším o zatraktivnění jinak nudné stránky vzorového vyhledávače.

4.8 Příklady

Pro testování vzorového vyhledávače bylo zvoleno několik menších webů, z nichž nejvhodnějšími k prezentaci dosažených výsledků jsou `http://www.jakpsatweb.cz/` a `http://www.ventilatory-kadlec.cz/`.

Testování výsledků dopadá v rámci možností hodnoticí funkce vyhledávače úspěšně. Projevují se nedostatky při hledání delších výrazů, kdy vyhledávač není schopen upřednostňovat stránky, na nichž se nachází výraz jako celek (fráze) a místo nich vybere stránku s častým výskytem některého nebo několika slov z hledaného výrazu. Správně funguje české skloňování i odstraňování diakritiky, což může být v českém národním prostředí hodnoceno jako vynikající funkce, i další nastavení vyhledávání.

Výpis vyhledávání zahrnuje pouze pět prvních položek výsledku a je formátován jako:

- poměrná váha objektu (mezera) jeho URL adresa

4.8.1 Případ 1 – „VE-450“

Vyhledání přesného typu produktu ve firemní prezentaci `http://www.ventilatory-kadlec.cz/` zcela neuspělo. Nejvýše byl ohodnocen rozcestník produktů, až na druhém místě konkrétní produktová stránka, kterou jsme hledali spíše.

- `282.0000 http://www.ventilatory-kadlec.cz/axv_ve.php`
- `123.5000 http://www.ventilatory-kadlec.cz/axv_ve-450.php`
- `104.0000 http://www.ventilatory-kadlec.cz/axv_ve-315.php`
- `97.0000 http://www.ventilatory-kadlec.cz/axv_ve-350.php`
- `90.0000 http://www.ventilatory-kadlec.cz/axv_ve-400.php`
- `...`

4.8.2 Případ 2 – „specifické použití“

Naopak vyhledání slovního spojení „specifické určení“, přestože v přesně takovémto tvaru se na stránkách vůbec nenachází, dopadlo dobře a byla nalezena očekávaná relevantní stránka.

- `164.5000 http://www.ventilatory-kadlec.cz/axv_spec.php`
- `18.5000 http://www.ventilatory-kadlec.cz/axv_info.php`
- `12.5000 http://www.ventilatory-kadlec.cz/axv_avet-315.php`
- `11.0000 http://www.ventilatory-kadlec.cz/axv_ve.php`
- `11.0000 http://www.ventilatory-kadlec.cz/axv_ve-315.php`
- `...`

4.8.3 Případ 3 – „Medvídek Pů“

Slovní spojení „Medvídek Pů“ je složeno z různě velkých písmen, je česky a v prvním pádě. Z webu `http://www.jakpsatweb.cz/` byly správně na prvních místech nalezeny dokumenty, kde se vyskytuje vícekrát i v jiných pádech a dále i jen jako „pů“:

- `24.0000 http://www.jakpsatweb.cz/weblog/archiv/2003-08.html`
- `20.0000 http://www.jakpsatweb.cz/clanky/jak-prekladate-site.html`
- `20.0000 http://www.jakpsatweb.cz/clanek/jak-prekladate-site.html`
- `17.5000 http://www.jakpsatweb.cz/weblog/archiv/2003-06.html`
- `16.0000 http://www.jakpsatweb.cz/weblog/archiv/200304.html`
- `...`

4.8.4 Případ 4 - „java script“

V tomto případě chceme vyhledat v rozsáhlém webu poměrně často se vyskytující slovo „javascript“, bohužel jsme výraz zadali chybně s mezerou jako „java script“. Přesto si vyhledávač s dotazem poradil a našel relevantní dokumenty. Na prvním a čtvrtém místě rozcestníky webu, na třetím a pátém odkaz na hlavní stránku části o JavaScriptu:

- 920.5000 http://www.jakpsatweb.cz/navigace/mapa-jakpsatweb-cz.html
- 451.5000 http://www.jakpsatweb.cz/css/behavior/prvni.gif
- 345.5000 http://www.jakpsatweb.cz/javascript/priklady/index.html
- 216.0000 http://www.jakpsatweb.cz/enc/encyklopedie.html
- 177.0000 http://www.jakpsatweb.cz/javascript/
- ...

5 Závěr

V této práci byly popsány způsoby vyhledávání v obsahu webu pomocí katalogových, full-textových a metavyhledávačů, popsány nejčastější algoritmy používané při full-textovém vyhledávání a navržen, implementován a otestován vzorový full-textový vyhledávač. Datový nosič s tímto projektem je přiložen k práci a je její nedílnou součástí.

Možné pokračování této práce vidím v hlubším poznání a srovnání full-textových vyhledávacích algoritmů na reálných vzorcích webu a v rozšíření vzorové implementace vyhledávače o vlastnosti těchto algoritmů, stejně jako začlenění dalších nadstandardních vlastností naznačených v kapitole *Rozšíření vyhledávačů* a to zejména v oblasti metavyhledávání, která je ve vodách českého Internetu zcela opominuta.

6 Slovníček

- **server** - obecné označení pro počítač (hardware) nebo proces (software), který poskytuje nějakou službu
- **klient** - obvykle program, který přistupuje k serveru
- **protokol** - soubor syntaktických a sémantických pravidel určujících formu komunikace
- **http** - Hyper Text Transfer Protocol - internetový protokol určený původně pro výměnu hypertextových dokumentů ve formátu HTML mezi serverem a klientem
- **ftp** – File Transfer Protocol – internetový protokol určený k přenosu souborů mezi serverem a klientem
- **www** - World Wide Web (WWW, také pouze zkráceně web) - označení pro soustavu propojených hypertextových dokumentů
- **html** - HyperText Markup Language - značkovací jazyk pro hypertext; jeden z jazyků pro vytváření stránek v systému WWW, který umožňuje publikaci stránek na Internetu
- **URL** - Uniform Resource Locator - řetězec znaků s definovanou strukturou sloužící k přesné specifikaci umístění zdrojů informací (ve smyslu dokument nebo služba) na Internetu
- **relevance** – většinou poměrná hodnota vyjadřující míru vyhovění výsledku vyhledávání jeho zadavateli

7 Literatura

[1] Pánek, K.: Jak pracuje metavyhledávač?
Dostupný na WWW: http://www.lupa.cz/clanky/jak-pracuje-metavyhledavac/

[2] Janovský, D.: Jak přibližně pracují vyhledávače
Dostupný na WWW: http://www.jakpsatweb.cz/vyhledavace.html

[3] Fishkin, R.: Search Engine Ranking Factors V2
Dostupný na WWW: http://www.seomoz.org/article/search-ranking-factors

[4] Houdek, A.: Způsoby hodnocení relevance vyhledaných dokumentů ve vyhledávacích strojích
Dostupný na WWW: http://www.ikaros.cz/node/1132

[5] Janovský, D.: Google PageRank
Dostupný na WWW: http://www.jakpsatweb.cz/seo/pagerank.html

[6] Page, L., Brin, S.: The PageRank Citation Ranking: Bringing Order to the Web
Dostupný na WWW: http://www.voelspriet2.nl/PageRank.pdf

[7] Wikipedia: HITS algorithm
Dostupný na WWW: http://en.wikipedia.org/wiki/HITS_algorithm

[8] Tsaparas, P.: Link Analysis Ranking
Dostupný na WWW: http://www.cs.helsinki.fi/u/tsaparas/publications/PhD.Thesis.ps

[9] Lempel, R.: Introduction to Link Structure Analysis
Dostupný na WWW: http://webcourse.cs.technion.ac.il/236620/Winter2006-2007/ho/WCFiles/lec9-moreLinkAnalysis.pdf

[10] Veřejná konference SEO
Dostupný na WWW: http://seo.nawebu.cz/

Všechny zdroje citovány z WWW dne 12.5.2007.

Příloha A: Struktura databáze

sites	vstupní body indexovaných webů
ID	
date	datum vložení vstupního bodu webu
entrypoint	zadaná URL adresa vstupního bodu webu
status	stav indexace (čeká, probíhá, hotova...)

objects	indexované objekty (soubory)
ID	
date	datum poslední indexace
changed	datum poslední změny
nextdate	datum příští indexace
url	relativní url stránky
site	odkaz ke kterému webu stránka patří
language	zkratka jazyka, ve kterém je (zřejmě) psána
size	poslední známá velikost objektu
lastsize	poslední známá nenulová velikost objektu
info	(rozšířený) typ objektu
weight	hodnocení objektu

words	unikátní slova
ID	
word	slovo
occurences	celkový počet výskytů slova

occurences	indexované výskyty slov
ID	
objects	příslušnost výskytu objektu
target	příslušnost odkazu objektu
word	slovo
pos_char	poloha slova na stránce ve znacích
pos_word	poloha slova na stránce ve slovech
pos_par	poloha slova na stránce v odstavcích
appearance	bitové pole příznaků obsahu v hodnocených elementech
weight	váha výskytu slova podle součtu hodnocených elementů

accuracy	předvypočtené relevance výskyt slova / objekt
ID	
word	slovo
objects	objekt
appearance	kopie pole appearance z tabulky occurences
weight1	součet hodnot všech výskytů slova
weight2	součet hodnot všech objektů, kde se slovo vyskytuje
count	počet výskytů slova v textu objektu
clicks	počet kliknutí na výsledek vyhledávání

languages	jazyky dostupné v rozhraní vyhledávače
ID	
name	název jazyka
short	zkratka jazyka
valid	platnost jazyka

strings	multijazykové texty
ID	
language	jazyk
name	název řetězce
text	text řetězce

users	uživatelské účty administračního rozhraní
ID	
name	jméno uživatele
login	přihlašovací jméno uživatele
password	šifrované heslo
permissions	výčet oprávnění uživatele
settings	řetězec uživatelského nastavení prostředí

Příloha B: Uživatelská příručka

Uživatelská příručka

IBP2007

Autor: Jiří Tužil, xtuzil00@stud.fit.vutbr.cz
Fakulta Informačních Technologií
Vysoké Učení Technické v Brně

Obsah

1 Úvod

IBP2007 je webová aplikace určená k vyhledávání dat v prostředí webu. Jejím primárním účelem není indexovat miliardy dokumentů celého Internetu, ale zaměřuje se na indexování jednotlivých samostatných webů (serverů), kde nemusí být žádoucí přítomnost cizího vyhledávacího nástroje. Příkladem za všechny může být firemní intranet, který nebyl od počátku navržen s možností vyhledávání, anebo pomocí jeho vestavěných nástrojů není možné vyhledat informaci ve všech jeho částech jediným dotazem.

IBP2007 umožňuje indexovat libovolné množství webů pouhým vložením URL adresy vstupního bodu a v nich poté vyhledávat stránky, obrázky, soubory a e-mailové adresy podle zadaného výrazu.

2 Koncepce

IBP2007 sestává ze tří částí:

- webového rozhraní pro správu vyhledávače
- skriptu pro indexování webu
- webového rozhraní pro vyhledávání a prezentaci výsledků

Více bude o každé z těchto částí řečeno v samostatné kapitole.

3 Požadavky

K uložení dat je používán databázový server MySQL ve verzi alespoň 4.1 s podporou tabulek InnoDB. Webová část aplikace vyžaduje ke svému běhu jakýkoliv HTTP server, který podporuje interpretaci skriptů jazyka PHP a samotný PHP interpretr ve verzi alespoň 4. Protože celá aplikace pracuje nativně s kódováním znaků UTF-8, je potřeba rozšíření PHP mbstrings.

Tyto požadavky splňuje většina dnešních webhostingových služeb, přesto se prosím nejdříve informujte u Vašeho provozovatele webhostingu na konkrétní podmínky a instalované verze produktů.

4 Instalace

4.1 Distribuce

IBP2007 je distribuována jako jediný zip či tar archiv. Můžete ho získat v publikaci bakalářské práce s názvem *Vyhledávání dat*, která bude či byla (či nebyla :) obhájena na FIT VUT v Brně v červnu 2007. Jeho obsah jednoduše rozbalte do adresáře přístupného Vašemu HTTP serveru.

4.2 Nastavení MySQL serveru

Máte-li již na MySQL serveru vytvořenou databázi (přidělenou např. správcem databázového serveru), naimportujte do ní SQL kód ze souboru `indexer.sql`, který se nachází v distribučním archivu. V opačném případě jste si sami svým správcem a musíte nejprve vytvořit novou databázi a nového uživatele s oprávněním plného přístupu pouze k této databázi a poté do ní soubor `indexer.sql` importovat. V obou případech by nyní měly být vytvořeny všechny potřebné tabulky pro běh IBP2007.

Máte-li zálohu databáze z předchozí instalace (při upgradu nebo havárii databáze), importujte ji do databáze nyní.

4.3 Úprava konfiguračních souborů

V souboru `/config/db.inc.php` opravte parametry pro připojení k databázi podle údajů, které jste obdrželi od správce serveru. Jsou to `DB_HOST` (adresa MySQL serveru), `DB_USER` (jméno uživatele pro přihlášení k MySQL serveru), `DB_PASS` (heslo pro přihlášení k MySQL serveru) a `DB_NAME` (název databáze).

Další konfigurační soubory, ve kterých byste mohli chtít něco změnit jsou `/config/config.inc.php` pro globální parametry, `/admin/_config.inc.php` pro parametry administračního rozhraní a `/spider/_config.inc.php` pro parametry indexovacího skriptu. Dokumentaci k jednotlivým parametrům najdete uvnitř každého konfiguračního souboru.

4.4 Nastavení automatické reindexace

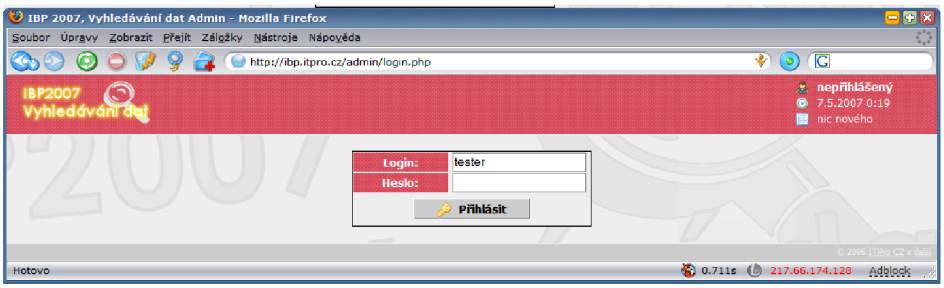
Posledním krokem instalace je nastavení automatické reindexace. Skript `/spider/reindex` (resp. `/spider/reindex.bat` pro operační systém Microsoft Windows) po spuštění zkontroluje objekty v databázi a reindexuje ty, které jsou zastaralé nebo naopak úplně nové. Reindexaci je možné spouštět z administračního rozhraní ručně, pohodlnější a spolehlivější ovšem je nastavit spuštění tohoto skriptu do plánovače úloh Vašeho operačního systému např. jednou denně, aby byly zastaralé objekty reindexovány v potřebný čas bez nutnosti Vašeho dalšího zásahu.

5 Webové rozhraní pro správu vyhledávače

Webové rozhraní pro správu vyhledávače (nebo jednodušeji administrační rozhraní) je pro správce systému hlavním bodem aplikace. Předpokládejme, že adresa vyhledávače IBP2007 na Vašem webovém serveru je `http://ibp.itpro.cz/`. Administrační rozhraní se potom nachází v podadresáři `admin/`, tzn. na adrese `http://ibp.itpro.cz/admin/`.

5.1 Přihlášení a odhlášení

První stránkou, kterou po zadání adresy administračního rozhraní do webového prohlížeče vidíte, je přihlašovací dialog. Protože ještě nemáte vytvořené žádné uživatelské účty, použijte přihlašovací jméno „tester“ a heslo „tester“ testovacího uživatele (obojí bez uvozovek).



Nyní jste přihlášen jako uživatel „tester“. Až skončíte svou práci v administračním rozhraní, nezapomeňte se stejně jako vždy odhlásit kliknutím na volbu *Odhlásit* v menu a zavřít okno prohlížeče. Přestože administrační rozhraní Vás po čase neaktivity (definovaném konstantou `LOGIN_TIMEOUT` v souboru `/admin/_config.inc.php`) odhlásí samo, nemusí to být vždy dostatečně rychle na to, aby se nemohlo něco nepříjemného přihodit.

5.2 Navigace

Práce s administračním rozhraním je intuitivní a snadno se v něm zorientujete. Nejste-li si jisti, co která volba znamená, podržte nad ní ukazatel myši – většina voleb, tlačítek nebo odkazů, je-li to předmětné, zobrazí rozšířený popis své akce.

Může se stát, že v dalších kapitolách nebude Vaše stránka v některých aspektech odpovídat popisu. Bude to pravděpodobně způsobeno omezením Vašich práv, která některé odkazy,

tlačítka nebo celé stránky znepřístupní. Není-li to záměr správce, dozvíte se, jak vytvořit nový uživatelský účet se všemi oprávněními, v kapitole Správa uživatelských účtů.

Pokud narazíte na zajímavost, podivnost nebo problém, nebojte se experimentovat nyní – dříve, než budete mít databázi plnou důležitých dat!

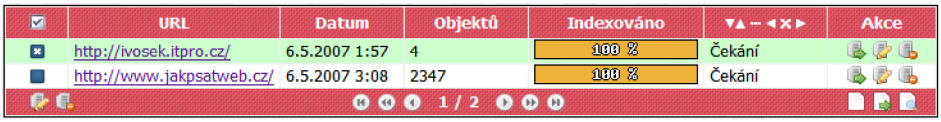
5.2.1 Menu

Podstatnou částí je černé roletové menu v hlavičce stránky, které funguje jako rozcestník k dalším stránkám popisovaným v následujících kapitolách. Jeho malou zvláštností může být prolinkování už od vodorovné lišty, což přispívá k rychlejší navigaci. V roletkách jsou pak většinou na prvních místech přímo odkazy na akce, pod oddělovačem pak jen na přehledy dat.

Aby roletové menu fungovalo správně, musíte mít ve Vašem webovém prohlížeči povolené spouštění JavaScriptu.

5.2.2 Tabulky

Pro administrační rozhraní jsou typické tabulky – uvidíte je téměř na každé stránce prezentující různé druhy informací.



Jejich výhodou je univerzálnost a velká modifikovatelnost – najetím na záhlaví každého sloupce se objeví několik ikon, s jejichž pomocí můžete:

-  – seřadit tabulku podle sloupce sestupně
-  – seřadit tabulku podle sloupce vzestupně
- – vypnout/zapnout zkracování obsahu buněk ve sloupci
-  – přesunout sloupec vlevo
-  – skrýt sloupec
-  – přesunout sloupec vpravo

Řazení tabulky má paměť, to znamená, že můžete řadit nejprve podle jednoho sloupce a poté podle jiného. Zkracování textu v buňkách je velmi užitečné ve chvíli, kdy se stránka doslova „rozplizne“ přes tři šířky obrazovky, třeba kvůli dlouhým odkazům v buňkách – a velmi

neužitečné ve chvíli, kdy na ně chcete kliknout. Počet znaků, na které se obsah buňky zkracuje, udává konstanta `MAXCELLLEN` definovaná v souboru `/admin/_config.inc.php`. Přesouvání a skrývání sloupců je pak čistě věcí osobní preference. Nutno podotknout, že vrátit zpět skrytý sloupec je možné jen smazáním Vašeho uživatelského nastavení (viz kapitolu Smazání uživatelského nastavení).

Ke každému záznamu (řádku) přísluší ikony akcí, které s ním lze provést. První ze zobrazených akcí je možné vybrat také dvojklikem kdekoliv na řádku (většinou to bývá detail záznamu a nikdy to není smazání). V některých tabulkách je možné označit více záznamů najednou, jednoduše jedním kliknutím kdekoliv na řádek. V levé části zápatí tabulky se pak objeví ikony akcí, které je možné provést s celým výběrem. Potřebujete-li vybrat všechny zobrazené záznamy najednou, nemusíte klikat na všechny řádky, stačí kliknout na obrázek zaškrtávátka v záhlaví prvního sloupce a výběr všech záznamů se obrátí.

Někdy je záznamů více, než se vejde na jedu stránku tabulky. To poznáte podle zobrazeného číselníku a šesti tlačítek pro navigaci mezi stránkami tabulky uprostřed jejího zápatí:

-  – přejít na první stránku
-  – přejít o polovinu stránek zpět
-  – přejít na předchozí stránku
- číslo stránky / celkový počet stránek
-  – přejít na následující stránku
-  – přejít o polovinu stránek vpřed
-  – přejít na poslední stránku

Záznamů ale může být tolik, že je neustálé přecházení mezi stránkami na obtíž. Proto jsou v pravé části zápatí tabulky ještě další možnosti:

- zobrazit všechny záznamy tabulky na jednu stránku
-  vrátit zobrazení záznamů na více stránek
-  přejít přímo na zvolené číslo stránky
-  zobrazit pouze záznamy obsahující zadaný výraz

Zobrazení všech položek na jednu stránku může dát Vašemu prohlížeči dost práce – používejte tedy toto zobrazení raději jen v případě menšího počtu záznamů v tabulce, v řádu stovek. Mnohem výhodnější je zobrazit jen ty záznamy, které opravdu chcete vidět, definováním filtru výběru.

Filtrovací výraz sestává ze slov spojených operátory `AND a OR` (na velikosti písmen záleží), jejichž prioritu můžete potlačit závorkami `( a )`. Slovo pak může být bezprostředně uvozeno symbolem `!`, pokud máte na mysli negaci jeho výskytu.

Příklad:

- `auto AND (modré OR žluté)` – vybere záznamy, které obsahují „auto“ a zároveň obsahují „modré“ nebo „žluté“ (nebo oboje)
- `motorka OR auto AND !sedan` – vybere záznamy, které obsahují „motorka“ nebo obsahují „auto“, ale s ním zároveň neobsahují „sedan“ (tzn. že „sedan“ spolu s „motorka“ je v pořádku)

Zrušení filtru provedete nastavením prázdného filtrovacího výrazu.

5.2.3 Dialogy

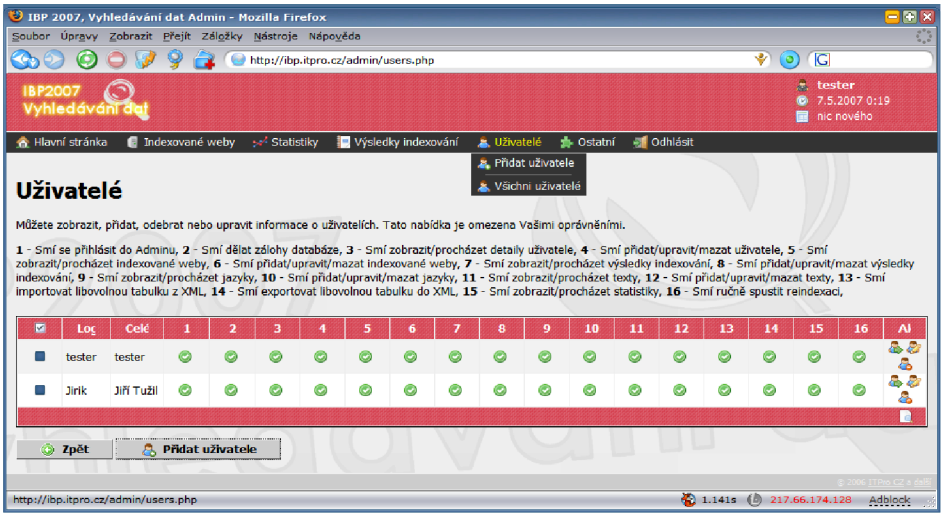
Typizovanou součástí administračního rozhraní jsou také informační dialogy. Jejich význam je zřejmý na první pohled podle jejich podbarvení:

- červené – oznamují chybové hlášení nebo důležitý dotaz
- zelené – informují o úspěšném provedení akce
- žluté – informují o jiné neobyčejně zajímavé události

Dialog (s výjimkou dotazu) se ve výchozím nastavení chová tak, že zobrazí informační hlášení, počká dobu určenou jeho délkou a poté automaticky přejde zpět nebo na další stránku. Pokud nechcete tak dlouho čekat, potvrzovat každý dialog tlačítkem, nebo je pro Vás naopak mizení dialogů příliš rychlé, můžete potlačit toto chování změnou konstanty `DIALOG_TIMEOUT` v konfiguračním souboru `/admin/_config.inc.php`. Nastavením konstanty na nulovou hodnotu dialogy prakticky zrušíte.

5.3 Správa uživatelských účtů

Administrační rozhraní může používat zároveň více uživatelů. Aby je bylo možné identifikovat a přidělit jim určitá oprávnění, musíte definovat jejich identity – uživatelská jména, hesla a další vlastnosti. Na stránku správy uživatelských účtů se dostanete kliknutím na položku *Uživatelé* v hlavním menu.



Jednotlivé uživatelské účty nyní vidíte v tabulce spolu s jednotlivými povolenými nebo zamítnutými oprávněními a možnými akcemi k uživatelskému účtu. Každá akce uživatele v administračním rozhraní se řídí příslušným oprávněním. Při zakládání nového uživatelského účtu tedy můžete přesně rozhodnout o tom, které akce uživateli povolíte a které zamítnete. Návštěvníkovi např. povolíte pouze přihlášení a prohlížení záznamů, operátorovi zálohování pouze přihlášení a zálohování, někomu, kdo Vás zrovna naštval, jednoduše zamítnete přihlášení atp. Zobrazení detailů, úpravu nebo smazání uživatelského účtu provedete vždy kliknutím na příslušnou ikonu na řádku uživatele.

Dostupné akce na této stránce:

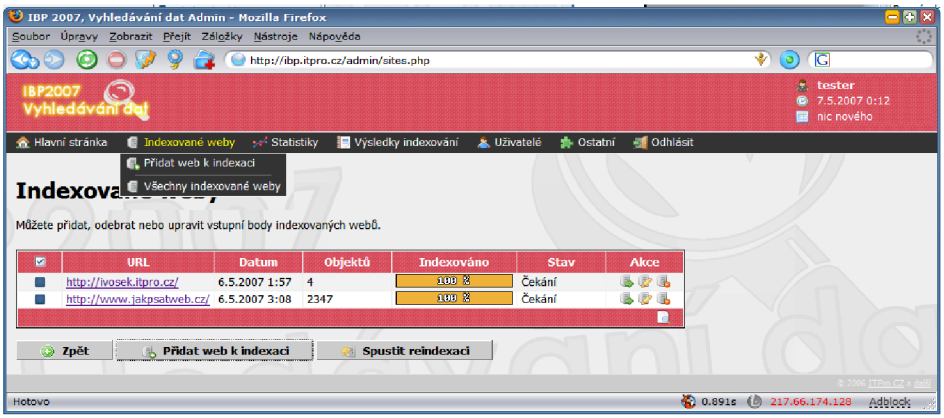
- – přidat uživatele
- – zobrazit detail uživatele
- – upravit uživatele
- – smazat uživatele

Nyní vytvořte účet pro každého uživatele, který bude administrační rozhraní používat, a nastavte mu potřebná práva. Nezapomeňte vytvořit svůj vlastní účet s úplnými právy!

Pokud už jste vytvořili všechny potřebné uživatelské účty, smažte testovacího uživatele „tester“ a odhlaste se z administračního rozhraní. Od této chvíle se už budete přihlašovat výhradně svým vlastním uživatelským jménem a heslem.

5.4 Správa indexovaných webů

Přidávat, sledovat, prohlížet, upravovat a mazat indexované weby můžete na této stránce.



Přehled indexovaných webů a jejich vstupních bodů vidíte v tabulce spolu s datem vložení, počtem objektů příslušejících indexovanému webu a procentuálnímu a slovnímu vyjádření stavu indexace (pro bližší popis stavů indexace viz kapitolu Skript pro indexování webu). Vstupní bod každého webu můžete odebrat nebo upravit. Mějte na paměti, že i v druhém případě budou vymazána veškerá indexovaná data tohoto webu a bude muset být reindexován od nově zadaného vstupního bodu.

Z této stránky můžete také ručně spustit reindexaci nových nebo zastaralých objektů, pokud nemáte nastavenou automatickou reindexaci (viz kapitolu Nastavení automatické reindexace). Reindexace bude spuštěna na pozadí, její aktuální stav uvidíte v tabulce indexovaných webů a po dokončení si budete moci výsledek reindexace prohlédnout na stránce *Výsledky indexovaní*.

Dostupné akce na této stránce:

- – přidat web k indexaci
- – zobrazit detail indexovaného webu
- – upravit vstupní bod indexovaného webu
- – smazat indexovaný web včetně všech příslušejících dat
- – manuálně spustit kontrolu a reindexaci databáze

Nyní zkuste přidat nový web k indexaci zadáním URL jeho vstupního bodu (např. `http://www.itpro.cz/`) a stiskněte tlačítko *Spustit reindexaci*. Příštích několik minut či

hodin (v závislosti na rozsahu indexovaného webu) můžete v tabulce sledovat postup indexování.

5.4.1 Detail indexovaného webu

Dvojklikem na řádek v tabulce indexovaných webů přejdete na stránku *Detailu indexovaného webu*. Mimo souhrnných informací o indexovaném webu je zde také seznam všech objektů, které pod vybraný web patří. U každého vidíte jeho rozpoznaný typ, poslední známou velikost v bajtech a tři data – přidání, poslední indexace a příští reindexace.

Je-li datum příští reindexace objektu příliš vzdálené nebo víte, že již nyní objekt neodpovídá stavu, ve kterém byl k datu poslední indexace, klikněte na ikonu akce *Reindexovat objekt*. Při příštím spuštění skriptu pro indexování webu bude tento objekt reindexován. Dvojklikem na řádek objektu se můžete dále dozvědět podrobné informace o indexovaném objektu.

Dostupné akce na této stránce:

-  – zobrazit detail objektu
-  – reindexovat objekt
-  – upravit vstupní bod indexovaného webu
-  – smazat indexovaný web včetně všech příslušejících dat

5.4.2 Detail indexovaného objektu

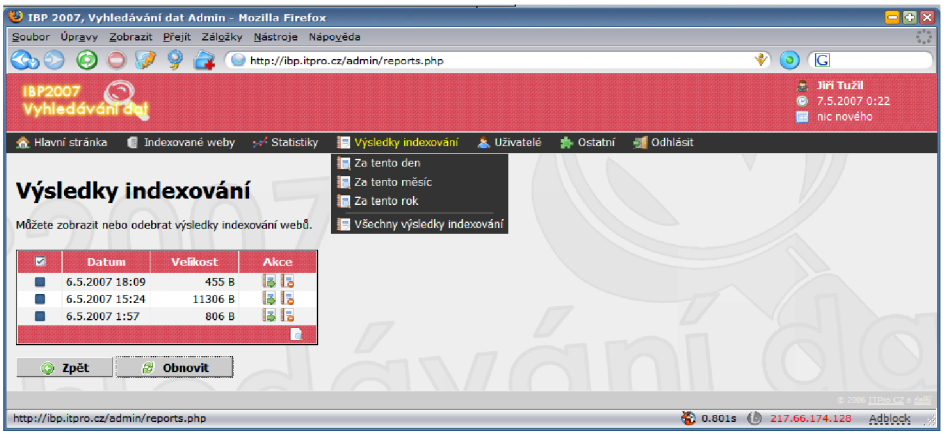
Na stránce *Detailu indexovaného objektu* vidíte všechny vztahy, které jsou pro daný objekt zaznamenány. Jsou to jednak odkazy na slova, která objekt popisují a jejich vzájemná poloha, ale také slova, která odkazují na jiné objekty a slova, která naopak odkazují z jiných objektů na tento. Právě podle těchto informací je vyhledávačem určována relevance objektu k zadanému vyhledávacímu výrazu.

Dostupné akce na této stránce:

-  – reindexovat objekt

5.5 Výsledky indexování

Indexace webů probíhá na pozadí. Je proto potřeba mít způsob, jak zpřístupnit její výstup – výsledek.



Každé spuštění skriptu pro indexování webu, automatické i ruční, po svém dokončení uloží datum a čas dokončení a výsledek své práce do tabulky na stránce *Výsledky indexování*, odkud si je můžete kdykoliv prohlédnout nebo je smazat. Více o výsledcích indexování se dozvíte v kapitole Skript pro indexování webu.

Dostupné akce na této stránce:

- – zobrazit detail výsledku indexování
- – smazat výsledek indexování

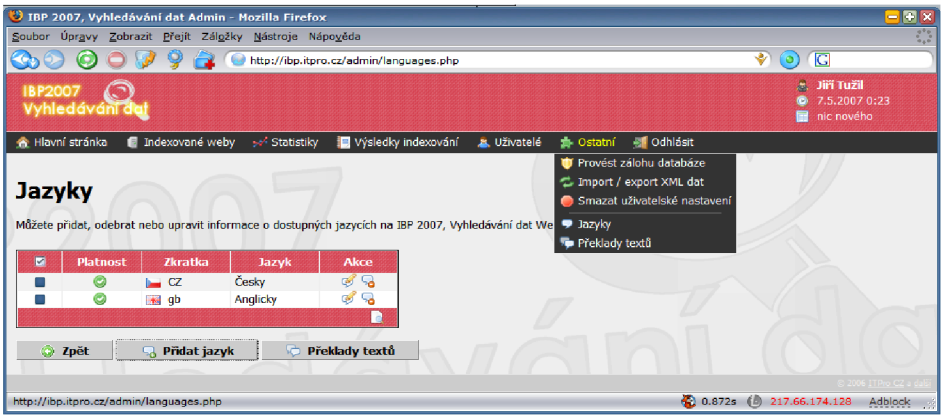
5.6 Statistiky

V této verzi IBP2007 nejsou statistiky nad indexovanými daty dostupné.

5.7 Správa jazyků, překlady textů

Protože vyhledávač umožňuje indexovat webové stránky v mnoha jazycích, je vhodné mít i vstupní stránku vyhledávače dostupnou ve více jazykových variantách. Dostupné jazyky a překlady textů do těch to jazyků můžete spravovat ze stránek *Jazyky* a *Překlady textů*, na které se dostanete z menu *Ostatní*.

Při přidávání nového jazyku nastavujete jeho platnost, název a zkratku. Platnost jazyka povolte až ve chvíli, kdy budete mít pro tento jazyk připravené všechny překlady. Zkratku jazyka uvádějte v mezinárodně platném dvoupísmenném formátu. Automaticky se podle ní přiřazuje ikona vlajky k jazyku a automatické zvolení jazyka při první návštěvě vyhledávacího rozhraní uživatelem.



Dostupné akce na stránce *Jazyky*:

- – přidat jazyk
- – upravit jazyk
- – smazat jazyk
- – přejít k překladům textů

Každý textový řetězec ve vyhledávacím rozhraní lze přeložit do dalších jazyků. Použijte ikonu akce *Upravit* u příslušného překladu a vyplňte překlady ve všech požadovaných jazycích.

Dostupné akce na stránce *Překlady textů*:

- – přidat překlad
- – upravit překlad
- – smazat překlad

Ve výchozí distribuci jsou dostupné překlady v českém a anglickém jazyce. Potřebujete-li více dostupných jazyků, přidejte je jako neplatné, doplňte všechny překlady a poté jazyky aktivujte.

5.8 Zálohování, import, export

Zálohování je velmi podstatnou částí každé činnosti, při které je něco hodnotného vytvářeno. Přestože data získaná indexací můžete znovu kdykoliv obdržet novou indexací, uložené vstupní body, uživatelské účty, jazyky, překlady textů a další informace, které jste sami zadali, se při havárii databáze již nemusí podařit znovu získat. Proto zálohujte, zálohujte, zálohujte.

Administrační rozhraní nabízí dva druhy zálohování – *Celkovou zálohu databáze* a *Export do XML*. Obě volby najdete v menu *Ostatní*. Liší se od sebe rozsahem a formátem výstupu. Nejste-li si jisti, že budete výstup ve formátu XML k něčemu dalšímu potřebovat, používejte *Celkovou zálohu databáze*, která Vám po kliknutí nabídne uložit aktuální stav dat na Váš počítač.

Dostupné akce pro zálohování:

-  – Provést zálohu databáze
-  – Import / export XML dat

Obnovu databáze v případě havárie provedete importováním uloženého záložního souboru do nově založené a připravené databáze podle kapitoly Nastavení MySQL serveru.

5.9 Smazání uživatelského nastavení

Administrační rozhraní si pamatuje všechno Vaše uživatelské nastavení, které v tabulkách a stránkách provádíte, pro příští přihlášení. Může se stát, že budete chtít své uživatelské nastavení zrušit a uvést vše do původního stavu jako při prvním přihlášení. V takovém případě zvolte položku *Smazat uživatelské nastavení* z menu *Ostatní* a potvrďte následující potvrzení.

6 Skript pro indexování webu

Skript pro indexování webu je tichý dříč, který prochází Vámi zadané webové servery, stahuje z nich stránky, obrázky a další soubory (souhrnně objekty), zjišťuje o nich informace a zapisuje je do databáze (indexuje), aby v nich později šlo rychle vyhledat požadované informace. Říká se mu také bot, robot, crawler nebo spider (angl. pavouk, protože „leze po síti“).

6.1 Chování skriptu pro indexování webu a jeho parametry

Soubor `reindex` s tímto skriptem a dalšími jemu příslušnými soubory najdete v adresáři `/spider/`. Tento soubor je spouštěn automaticky (nastavením v plánovači úloh operačního systému, viz kapitolu Nastavení automatické reindexace) nebo ručně (kliknutím na tlačítko *Spustit reindexaci* v administračním rozhraní) a provádí následující činnosti v tomto pořadí:

- ověří, že je spuštěn jen v jediné instanci, pokud ne, dále nepokračuje
- zkontroluje úplnost vstupních bodů indexovaných webů a chybějící doplní
- vyhledá nové nebo zastaralé objekty, které je nutné reindexovat a reindexuje je
- zkontroluje, zda při předchozí reindexaci nepřibyly nové objekty a pokud ano, vrací se na předchozí bod
- ze získaných dat připraví pro každou kombinaci slova a objektu odpovídající vzorek pro urychlení vyhledávacího procesu
- zapíše protokol o své činnosti do databáze

V každé z těchto fází se mění stav aktuálně indexovaného webu, který se zobrazuje v administračním rozhraní. Možné stavy indexace jsou:

- Čekání – žádný z objektů webu není v současné chvíli indexován
- Úklid databáze – probíhá kontrola vstupních bodů a závislostí mezi objekty
- Probíhá indexace – některý z objektů webu je právě (re)indexován
- Probíhá rekalkulace – přepočítávají se vztahy mezi objekty a slovy
- Indexováno, v údržbě – veškerá činnost byla úspěšně ukončena, databáze je v plně funkčním stavu a udržují se pouze zastaralé objekty

Pro každý web indexuje skript jen ty objekty, jejichž URL začíná stejně jako URL vstupního bodu tohoto webu. To znamená, že vede-li z webu se zadaným vstupním bodem `http://www.itpro.cz/` odkaz např. na stránku

`http://www.itpro.cz/produkty.html`, pak bude tato indexována také. Odkaz na stránku `http://www.seznam.cz/`, ale ani `http://produkty.itpro.cz/` už však pro web s uvedeným vstupním bodem indexován nebude.

Skript pro indexování webů uchovává ke každému indexovanému objektu také interval reindexace (tj. jak často má být objekt reindexován), který je stanovován podle častosti změny velikosti objektu. Výchozí, nejkratší a nejdelší interval reindexace jednoho objektu ve dnech je možné upravit změnou hodnoty konstant `PAGE_DEFAULT_REFRESH`, `PAGE_MIN_REFRESH` a `PAGE_MAX_REFRESH` v konfiguračním souboru `/spider/_config.inc.php`. Konstanta `PAGE_DIE` ve stejném souboru určuje dobu nedostupnosti objektu ve dnech, po které je objekt považován za zrušený a je odstraněn z databáze.

6.2 Podporované komunikační protokoly

Podporované typy komunikačních protokolů:

- `http://` – HTTP
- `https://` – zabezpečené HTTP
- `ftp://` – FTP
- `file://` – lokální souborový systém
- `mailto:` – fiktivní protokol pro indexaci e-mailových adres

6.3 Podporované formáty objektů

Podporované typy objektů:

- `text/html` – dokumenty v jazyce HTML / XHTML
- `text/xml` – dokumenty v jazyce XML
- `text/plain` – dokumenty v čistém nebo strukturovaném textu
- `image/png` – obrázky ve formátu PNG
- `image/gif` – obrázky ve formátu GIF
- `image/jpeg` – obrázky ve formátu JPEG
- `address/email` – e-mailové adresy

Poslední typ ve výčtu není standardní (MIME) typ jako ostatní. Byl navržen pro usnadnění práce s indexací e-mailových adres.

Všechny textové typy obsahují jako rozšířený parametr `charset`, který specifikuje použité kódování textu stránky. Obrázkové typy pak obsahují jako rozšířené parametry `width` a `height`, které specifikují jejich šířku a výšku v pixelech.

Celý rozšířený typ objektu pak vypadá například takto:

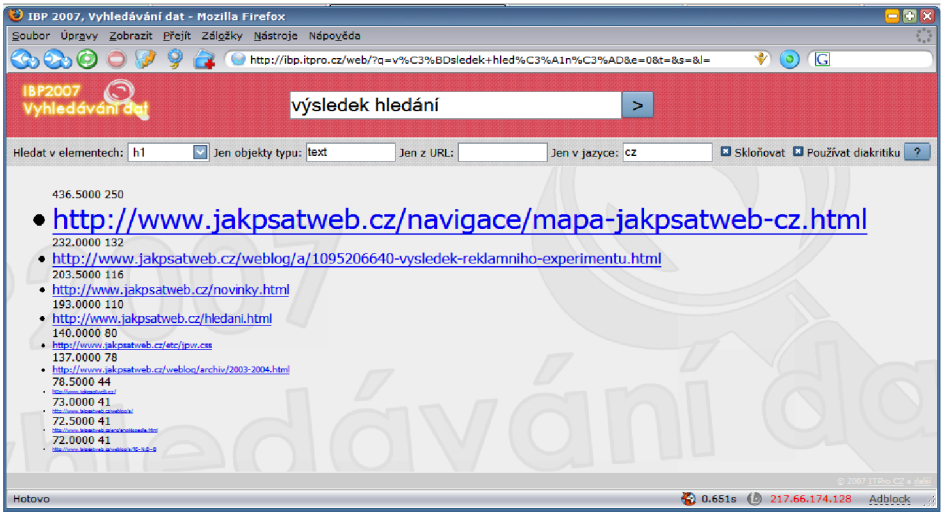
- `text/html; charset=iso-8859-2` – označuje textový dokument v jazyce HTML v kódování ISO 8859-2
- `image/jpeg; width=48; height=48` – označuje obrázek ve formátu JPEG vysoký i široký 48 pixelů

I podle těchto rozšířených parametrů typu je možné objekty vyhledávat.

7 Webové rozhraní pro vyhledávání a prezentaci výsledků

Webové rozhraní pro vyhledávání a prezentaci výsledků uvidí každý, kdo bude vyhledávač IBP2007 používat. Vstup na jeho úvodní stránku je přímo z kořene adresy vašeho HTTP serveru, v našem případě tedy `http://ibp.itpro.cz/`.

Úvodní stránka neobsahuje nic víc než formulář pro zadání vyhledávaného výrazu a tlačítko pro spuštění vyhledávání. Poté, nejste-li s výsledkem vyhledávání spokojeni, je Vám nabídnuta lišta s rozšířeným nastavením a krátkým nápovědným textem.



Rozšířené nastavení zahrnuje:

- hledání pouze ve vybraných elementech
- hledání jen objektů daného typu
- hledání jen těch objektů, jejichž URL obsahuje zadaný řetězec
- hledání objektů ve specifikovaných jazycích
- možnost skloňování českých slov
- možnost doplnění nebo odstranění diakritiky

Z těchto možností je třeba přiblížit snad jen vyhledávání podle typu objektu: typ objektu je výběr z výčtu podporovaných typů objektů uvedených v kapitole Podporované typy objektů a to včetně rozšířených parametrů; a jazyka: ten udává sama webová stránka a je to dvoupísmenné mezinárodně platné označení jazyka. V obou případech můžete uvést více

možností oddělených mezerou. Do výsledku budou zahrnuty jen takové objekty, které vyhovují všem zadaným parametrům.

Skloňování českých slov při vyhledávání bere v úvahu i všechny varianty slov v hledaném výrazu a to jak v jednotném, tak i množném čísle. Doplnění nebo odstranění diakritiky hledá i slova bez nebo naopak s háčky a čárkami, podle toho, jak byla slova zadána ve vyhledávacím výrazu.

Samotný vyhledávací výraz sestává ze slov oddělených mezerou. Na velikosti písmen nezáleží. Do výsledku vyhledávání budou zahrnuty ty objekty, které budou nejlépe odpovídat kombinaci slov v zadaném vyhledávacím výrazu. Další verze vyhledávače budou rozšířeny i o operátory `AND`, `OR`, `NOT`, `NEAR` ad., pro což nebude třeba znovu reindexovat zadané weby.

Prezentace výsledků je zcela přímá – na deseti řádcích jsou uvedeny odkazy na objekty v tom pořadí a takovou poměrnou velikostí písma, jakou objekty vyhověly vyhledávacímu výrazu. Kliknutím na odkaz z výsledků vyhledávání se zároveň zvyšuje jeho pravděpodobnost na vyšší pozici ve výsledku vyhledávání při příštím stejném nebo podobném vyhledávacím výrazu.

8 Potíže, problémy a známé chyby

Za krátkou dobu vývoje vyhledávače byly objeveny nesnáze, které nebylo možné do termínu vydání opravit. O těchto chybách a nedokonalostech víme a v příštích verzích budou postupně odstraňovány:

- Skript pro vyhledávání se může na některých webech zacyklit a indexování nikdy nedokončit. V tom případě je potřeba ho ručně „zabít“, označit stránky, které zacyklení způsobily za „navždy indexované“ a spustit reindexaci znovu. Toto ale není uživatelsky přívětivá operace.
- Vyhledávací výraz je příliš jednoduchý a nedovoluje přesnější vyhledávaní.

Přesto věříme, že budete s funkčností vyhledávače spokojeni a pokud ne, nebudete o tom příliš mluvit :)