Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky
Vyšší odborná škola informačních služeb v Praze

David Möhwald

Výběr vhodného search enginu pro společnost LMC s.r.o.

Bakalářská práce

2007

Prohlášení

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

V Praze dne 30. 11 .2007 ....

podpis

Abstrakt:

Cílem této práce je doporučit optimální řešení search enginu pro společnost LMC s.r.o. (dále jen LMC), která je provozovatelem pracovních serverů.

Dílčí cíle:

- Popsat současné využití search enginu ve společnosti LMC.
- Představit principy současných web search enginů.
- Provést analýzu současných řešení na trhu a vybrat taková, která jsou vhodná pro vyhledávání na pracovních serverech.
- Zhodnotit jednotlivé varianty search enginů na základě předem stanovených metrik a stanovit optimální řešení pro společnost LMC.

Text práce je rozdělen do 7 částí.

První část práce se zabývá představením společnosti LMC a popisem různých typů vyhledávání, které ve svých službách nabízí.

Druhá část popisuje požadavky, které odrážejí představu společnosti LMC vzhledem k předpokládanému vývoji.

Třetí část je zaměřena na principy fungování současných search enginů, jejich architekturu, analýzu vstupních dat, tvorbu indexů a následné vyhodnocení vstupních dotazů.

Obsahem čtvrté části je představit vhodné varianty řešení search enginů dostupných na trhu, a to jak v komerční sféře, tak ve formě open source projektů.

V páté části práce jsou definovány metriky pro vyhodnocení jednotlivých variant.

Šestá část představuje úvod do problematiky porovnávání výkonnosti jednotlivých variant search enginů.

Poslední sedmá část se zabývá vyhodnocením jednotlivých řešení pomocí stanovených metrik a doporučení vhodného řešení pro služby nabízené společností LMC.

Klíčová slova

Index, Search engine, Metrika, E-recruitment, Outsourcing, Algoritmus.

Abstract:

The main objective of this graduation thesis is a recommendation of an optimal search engine solution for the LMC company – the leading provider on the Internet labor market.

The list of all objectives is as follows:

- To describe existing search engines working in the LMC company.
- To show how the recent web search engines work in general.
- To analyze the present situation of search engine solutions and identify the most suitable ones for the Internet job search portals.
- To evaluate each solution by prior defined metrics and to suggest an optimal solution for the LMC company

This graduation thesis consists of seven parts.

The first part indroduces the LMC company and their services.

The second part presents all the company requirements for future development.

The third part describes functioning and architecture of current search engines, index creation, document analysis and incoming queries processing.

The fourth part presents several search engines variants that are available on commercial market as well as an open source projects.

The fifth part defines the metrics for the valuation of the variants.

The sixth part compares the search engine performance of defined variants.

The last part ranks each variant and proposes an optimal solution for the LMC company.

Key words

Index, Search engine, Metrics, E-recruitment, Outsourcing, Algorithm.

Obsah:

1 Úvod

S přibývajícím množstvím informací, které jsou dostupné uživatelům využívajícím internet, rostou také požadavky na vyhledávače. Především jde o to, jak z tohoto množství informací vybrat ty, které jsou pro uživatele nejvíce relevantní, vrátit je na výstup v co nejkratším čase a zároveň poskytnout jednoduché, ale intuitivní uživatelské rozhraní, které by uživatele vedlo k zadávání co nejvíce přesných dotazů.

Tato práce je zaměřena především na vyhledávání v oblasti **e-recruitmentu** neboli získávání pracovníků pomocí elektronických sítí. Služby tohoto typu jsou dnes již nepostradatelnou součástí každodenní práce personalistů ve většině středně velkých a velkých firem a představují jeden z nejefektivnějších zdrojů vhodných kandidátů.

Na opačné straně oproti nabídce firem přichází každý den tisíce uživatelů internetu s požadavky na vyhledání nového nebo pouze vhodnějšího zaměstnání. Místo, kde se střetává nabídka pracovních míst s poptávkou ze strany uchazečů o zaměstnání lze nazvat **elektronický trh práce**. Na elektronický trh práce v ČR lze přistoupit přes desítky zprostředkovatelů, kteří na internetu provozují **pracovní servery.** Největším provozovatelem pracovních serverů v ČR je firma LMC s.r.o., která bude blíže představena v následující kapitole.

Provozovat v dnešní době vlastní pracovní server není příliš náročné a svou složitostí se dá přirovnat například k jednoduchému bazaru nebo elektronickému obchodu. To však platí pouze v případě, kdy se detailněji nevěnujete vývoji na elektronickém trhu práce a nenaslouchátе požadavkům ze strany firem a uchazečů. Vzhledem k tomu, že jsem měl možnost sledovat v průběhu několika let vývoj v této oblasti, mohu říci, že komplexnost a složitost dnešních služeb, se kterými je možné se na elektronickém trhu práce setkat, je přirovnatelná k několika rokům intenzivního vývoje v týmu desítek lidí. Jako příklad služeb lze uvést kompletní workflow náborového procesu, napojení na tisková média nebo **různé typy relevantních vyhledávání**.

Na služby spojené s vyhledáváním je také zaměřena tato práce, která si klade za cíl představit možná řešení vyhledávacích strojů (search enginů) a doporučit optimální řešení pro jejich implementaci, se kterou se lze setkat na pracovních serverech provozovaných společností LMC. Při výběru vhodného řešení budou zohledněny nové požadavky,

nedostatky současného řešení a stanovené metriky včetně variant, kdy je s**earch engine** provozovaný na vlastních serverech nebo využívá tzv. outsourcingu, který ponechává kompletní správu a běh aplikace na straně dodavatele.

1.1 Cíle práce

Cílem této práce je doporučit optimální řešení search enginu pro společnost LMC s.r.o. (dále jen LMC), která je provozovatelem pracovních serverů.

Dílčí cíle:

- Popsat současné využití search enginu ve společnosti LMC.
- Představit principy současných web search enginů.
- Provést analýzu současných řešení na trhu a vybrat taková, která jsou vhodná pro vyhledávání na pracovních serverech.
- Zhodnotit jednotlivé varianty search enginů na základě předem stanovených metrik a stanovit optimální řešení pro společnost LMC.

1.2 Obsah práce

Text práce je rozdělen do 7 částí.

První část práce se zabývá představením společnosti LMC a popisem různých typů vyhledávání, které ve svých službách nabízí.

Druhá část popisuje požadavky, které odrážejí představu společnosti LMC vzhledem k předpokládanému vývoji.

Třetí část je zaměřena na principy fungování současných search enginů, jejich architekturu, analýzu vstupních dat, tvorbu indexů a následné vyhodnocení vstupních dotazů.

Obsahem čtvrté části je představit vhodné varianty řešení search enginů dostupných na trhu a to jak v komerční sféře, tak ve formě open source projektů.

V páté části práce jsou definovány metriky pro vyhodnocení jednotlivých variant.

Šestá část představuje úvod do problematiky porovnávání výkonnosti jednotlivých variant search enginů.

Poslední sedmá část se zabývá vyhodnocením jednotlivých řešení pomocí stanovených metrik a doporučení vhodného řešení pro služby nabízené společností LMC.

2 Popis současného řešení

2.1 Charakteristika společnosti LMC a nabízených služeb

Společnost LMC působí na trhu od roku 1996, kdy její zakladatel Ing. Libor Malý začal publikovat první pracovní nabídky na serveru jobs.cz. Postupem času se poptávka na elektronickém trhu práce začala zvyšovat a server jobs.cz si začal získávat stále větší popularitu, a to především z řad studentů a odborné veřejnosti. V roce 2002 již měla společnost LMC asi 20 zaměstnanců a přichází na trh s novým pracovním portálem prace.cz s výstižným heslem „Práce pro všechny“. Nový portál byl zaměřen především na širší veřejnost s nižší kvalifikací. Zároveň měl však sloužit jako agregátor všech pracovních nabídek na českém internetu.

Dnes lze jednoznačně říci, že je společnost LMC se svými službami v rámci České republiky jedničkou ve svém oboru a dá se jí jen velmi náročně konkurovat. Od roku 2007 působí společnost LMC se svými službami také na slovenském trhu, kde provozuje servery topjobs.sk a praca.sk. Zároveň je LMC zakladatelem společenství ONREA, které si klade za cíl umožnit zaměstnavatelům publikovat pozice na různé pracovní servery v rámci Evropy. Ke konci roku 2007 pracuje v LMC více jak 200 lidí, kteří se starají o zákazníky, propagaci a neustálý vývoj nových služeb. Již od počátku byl kladen důraz nejen na služby pro firmy, ale také na služby pro uchazeče o zaměstnání, bez kterých by samotný elektronický trh nemohl existovat.

Pro uživatele pracovních serverů byla vytvořena webová aplikace s označením „LMC G2“, do které se uživatelé přihlašují přes veřejnou prezentační vrstvu (VPV). Tato aplikace umožňuje firmám v části, kterou budu pracovně nazývat „G2NAS“, především publikovat volné pozice, vyhledávat vhodné kandidáty a vést celý proces náborového procesu. Na druhé straně v části zvané „G2MUJ“ mohou uchazeči o zaměstnání publikovat své životopisy, aktivovat automatizované vyhledávání nabídek a vést si správu procesu hledání práce.

2.2 Využití search enginu v LMC

LMC provozuje v současné době několik search enginů, které jsou přizpůsobeny struktuře dat, nad kterou vyhledávání probíhá, použitými algoritmy a tím, zda-li je nutné výsledky zobrazit v reálném čase nebo je postačí zpracovat později.

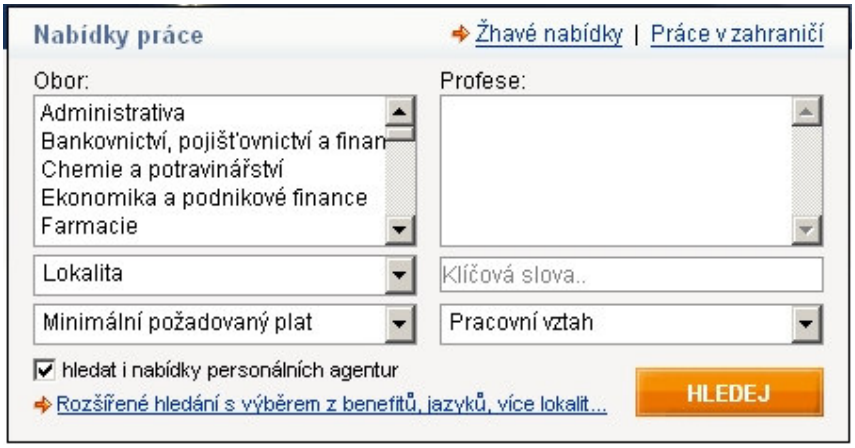
2.2.1 Online vyhledávání JD na VPV

Jedná se o vyhledávání v nabídce volných míst na VPV, které vrací nalezené dokumenty na základě zvolených kritérií v reálném čase. Z VPV je realizováno nejvíce dotazů na search engine, a proto je zde kladen velký důraz na výkonnost celého systému, rychlost nalezení a zobrazení výsledků.

Množinu vstupních kritérií lze dle způsobů vyhodnocování a použitých algoritmů rozdělit do 2 hlavních kategorií:

- hledání na klíčová slova,
- hledání na strukturovaná data.

Výstupní sada dokumentů je primárně přizpůsobena brandu, na kterém probíhá vyhledávání, kde se zohledňuje cílová skupina uživatelů, konkrétní obchodní model a způsob řazení výsledků. Režim vyhodnocování a řazení výstupu je možno zvolit striktní nebo relevantní.



Obrázek 2.2.1: Ukázka vyhledávacího formuláře (zdroj [Jobs.cz www])

2.2.2 Prohledávání brigád na VPV

Modul brigády představuje samostatný modul VPV, který prozatím nevyužívá možností žádného search enginu a poskytuje tak pouze omezené možnosti DB vyhledávání.

2.2.3 Online vyhledávání CV v G2NAS

V G2NAS mohou firmy v reálném čase prohledávat životopisy uchazečů o zaměstnání. Vyhledávání je řešeno obdobným způsobem jako na VPV, ale množina vstupních kritérií je přizpůsobena struktuře CV. V tomto typu vyhledávání není zatím umožněno volit způsob vyhodnocení, a proto probíhá ve striktním režimu. Oproti vyhledávání JD na VPV je tu menší náročnost na počet zpracovaných dotazů, ale je zde náročnější zpracování a dohledání jednotlivých dokumentů. To je způsobeno především tím, že uchazeč o zaměstnání může mít strukturované CV v několika jazykových mutacích a při odpovídání na volná pracovní místa má možnost pokaždé přiložit až 3 další binární přílohy.

2.2.4 Služba Distribution List v G2NAS

Tato služba umožňuje firmám oslovit vhodné uchazeče o zaměstnání na základě charakteristiky konkrétní pozice. Z pozice jsou vytaženy vstupní kritéria, která jsou použita při zavolání search enginu a pro vyhledání v hodných uchazečů. Následně je možné si zvolit, kolik uchazečů z nalezené množiny má být osloveno. Tato služba využívá shodného search enginu jako hledání CV v G2NAS.

2.2.5 Prohledávání správy náborového procesu v G2NAS

V modulu G2NAS mají firmy možnost si v rámci komplexní správy náborového procesu udržovat vlastní DB uchazečů, včetně historie provedených aktivit. Nad těmito daty není nyní zprovozněna žádná aplikace typu search engine a je zde proto využito pouze standardních možností DB vyhledávání.

2.2.6 Agenti

Jedná se o opakované spouštění vyhledávacího search enginu nad předem stanovenou množinou kritérií a na předem stanovenou dobu. Výsledné dokumenty se postupně zasílají uživateli a není tedy nutné, aby vyhodnocování probíhalo v reálném čase. Agent je reprezentován samostatným search enginem a je provozován ve dvou instancích, a to pro vyhledávání uchazečů a pro vyhledávání vhodných pozic. Agenti využívají režim relevantního vyhledávání s tím, že množina nalezených dokumentů je řazena dle relevance a je omezena předem stanoveným počtem. Pokud dojde k dosažení počtu nalezených dokumentů nebo expiraci doby platnosti agenta, potom je vyhledávání ukončeno.

2.3 Architektura a implementace

V LMC jsou preferovány tyto technologie:

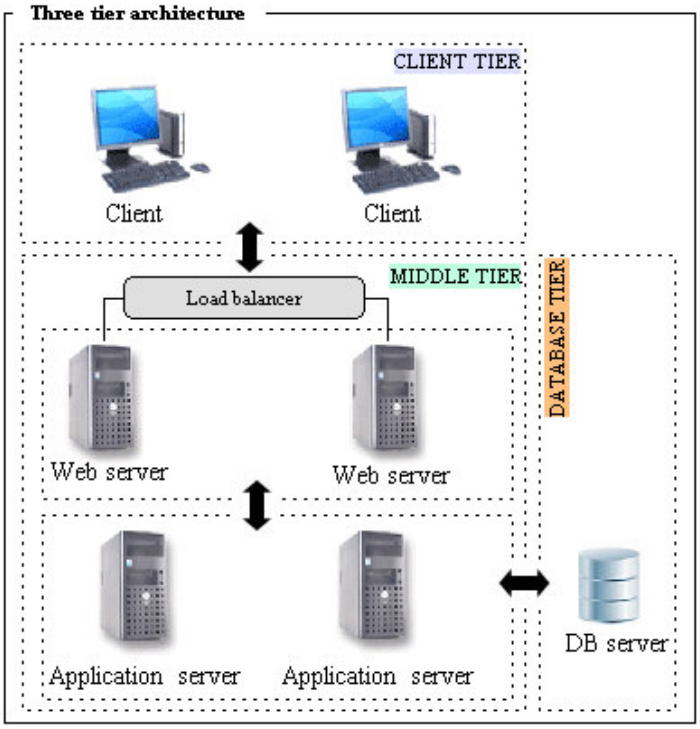
- OS LINUX, preferovaná distribuce CentOS,
- Java 1.4 a 1.5, Java Server Pages, aplikační server Tomcat,
- PHP 4.X a 5.X, webový server Apache,
- Databáze DB2, aplikační server WebSphere od IBM,
- Databáze PostgreSQL 7.X – 8.X.

2.3.1 3 vrstvá architektura

Vzhledem k vyšší výkonnosti a dosažení maximální dostupnosti služeb používá LMC vícevrstvou architekturu systému.

Požadavek uživatele směřuje z klientské vrstvy na webový server (Apache), který je uživateli schopen vracet přímo statický obsah (html soubory, obrázky apod.) a nebo předává požadavek na zpracování aplikačnímu serveru, který je schopen pracovat s různými skriptovacími jazyky (php, jsp, cgi apod.). Aplikační server při zpracování požadavku využívá dat uchovaných v databázi na odděleném databázovém serveru, který představuje poslední z vrstev. Mezi hlavní výhody vícevrstvé architektury patří možnost snadného přidávání nových serverů, které vede k rovnoměrnému rozdělení zátěže a zvýšení

spolehlivosti celého systému. Na druhé straně se zvyšuje náročnost na správu celého systému a na diskový prostor, který lineárně narůstá s množstvím použitých serverů, především díky nutnosti synchronizace dat.



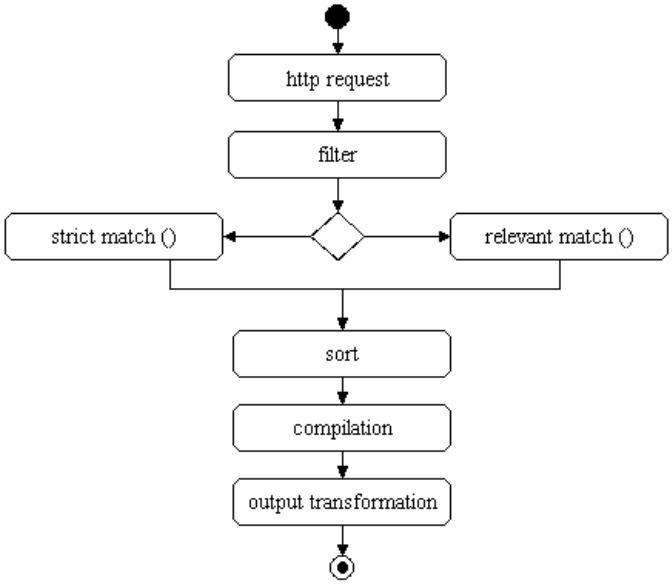
Obrázek 2.3.1 – Třívrstvá architektura systému

2.3.2 Aplikace Search

Aplikace Search využívaná v LMC na VPV je napsána v jazyce Java 1.5 a je spuštěna v několika instancích s využitím kontejneru Tomcat. Http požadavky na jednotlivé instance jsou rovnoměrně rozděleny pomocí load balanceru. Při spuštění nové instance aplikace dojde k prvotní inicializaci a načtení obsahu všech strukturovaných dat extrahovaných z dokumentů a uložených v DB (PostgeSQL) do cache, kterou v tomto případě představuje operační paměť. Do operační paměti se nenačítá pouze obsah DB určený pro fulltextové vyhledávání, pro které se nyní využívá rozšiřující modul zvaný T-search2. Proces vyhledání dokumentů je následující:

- Přijmutí a zpracování http požadavku.
- Vyhodnocení vstupního filtru.

- Vyhodnocení zda se jedná o relevantní nebo striktní fci.
- Spuštění fce a nalezení výsledkové množiny.
- Seřazení výsledkové množiny.
- Kompletace všech dat a řazení.



Obrázek 2.3.2 - Proces zpracování požadavku na aplikaci Search

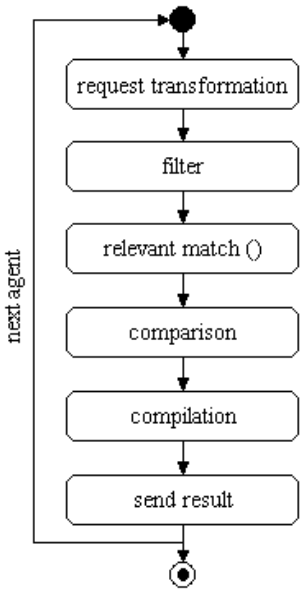
2.3.3 Aplikace Agenti

Aplikace Agenti je provozována na shodných technologiích jako aplikace Search. Rozdíl je pouze v použité DB (DB2 od IBM) a v principu dotazování a zpracování výsledků. Publikované dokumenty a množiny definovaných kritérií představující dotazy jsou odesílány ve formátu XML souborů pomocí webové služby do aplikace Agenti. Zde jsou datové struktury uloženy v DB a obdobně jako v aplikaci Search jsou převedeny na Java objekty a načteny do operační paměti. Fulltextové vyhledávání je realizováno pomocí databázového extenderu, který je rozšiřujícím modulem DB2.

Proces vyhledání dokumentů:

- Převedení množiny vstupních kritérií agenta na dotaz.
- Omezení množiny dokumentů pomocí fulltextu.
- Spuštění fce pro výpočet relevance a nalezení výsledkové množiny.

- Rozdílová analýza s odesláním pouze nových nebo více relevantních dokumentů.
- Předchozí postup vyjma prvního kroku se opakuje pro všechny agenty. Vyhodnocení však probíhá pouze nad množinou upravených nebo nových dokumentů.



Obrázek 2.3.3 – Proces vyhledávání jednotlivých agentů

2.3.4 Použité algoritmy

Pro hledání dle klíčových slov je ve všech uváděných řešeních využito pouze operátorů AND, OR nebo přesná fráze. Výsledkem je neuspořádaná množina záznamů obsahující zadaný výraz, bez určené relevance.

Níže uvedené algoritmy jsou využity pro většinu vstupních kritérií, které je možno definovat v dotazu.

Výsledná relevance jednoho kritéria RC je stanovena na základě váhy kritéria W vůči ostatním kritériím a je násobena mírou shody S. Celková váha dokumentu RT je potom stanovena jako součet relevancí všech kritérií a je normalizována normRT vůči maximální možné shodě ve všech zvolených kritériích.

$$RC = W \cdot S$$

$$RT = \sum_{i=1}^{n} RCi$$

$$normRT = \frac{\sum_{i=1}^{n} RCi}{\sum_{i=1}^{n} \max(RCi)}$$

Algoritmus pro porovnání dvou množin bez uspořádání

Lze si jej jednoduše představit jako porovnání dvou výčtových typů (např. číselník oborů nebo profesí). Stanovit míru shody S lze potom následovně:

- R=P – množina v kritériu R a dokumentu P je shodná, S=4
- R<P – kritérium je podmnožinou prvků dokumentu, S=3
- R>P – kritérium je nadmnožinou prvků dokumentu, S=3
- R>1<P – nepřesná shoda, S=2
- P=Null – prvky v dokumentu nebyly zadány, S=1
- R<>P – množiny se nerovnají, S=0

Algoritmus pro porovnání dvou množin s uspořádáním

Lze si představit jako množiny, kde lze považovat za shodu hodnoty, které jsou buď rovné a větší a nebo naopak rovné a menší. (např. plat nebo úroveň vzdělání).

- R=P – množina v kritériu a dokumentu je shodná, S=4
- R<=P – kritérium je menší nebo rovno prvkům v dokumentu, S=3
- R>=P – kritérium je větší nebo rovno prvkům v dokumentu, S=3
- P=Null – prvky v dokumentu nebyly zadány, S=1
- R<>P – množiny se nerovnají, S=0

Pro některá kritéria, např. pro vyhodnocení míry shody lokalit, existují samostatné algoritmy, jejichž popis však není primárním cílem této práce a vzhledem k jejich rozsahu při použití geografických systémů (GIS) a definic okolí bodů je zde nebudu uvádět.

2.4 Nedostatky současného řešení

Z kapitoly Popis současného řešení je zřejmé, že mezi hlavní nevýhody patří především rozdílnost jednotlivých implementací vyhledávání mezi aplikacemi. To je způsobeno použitím rozdílných technologií a strukturou dat, nad kterou se vyhledává, což má následující následky:

- Rozdílnost použitých typů databází způsobuje rozdílnou implementaci fulltextového vyhledávání, které **nepodporuje práci s český jazykem ani výpočet relevancí**.
- Aplikace využívají **rozdílné algoritmy a nastavení** pro vyhodnocení kritérií, a proto některé z nich stále nepodporují výpočet relevancí, i když by tento způsob vyhodnocování zlepšil kvalitu služeb.
- Z předchozích dvou bodů plynou i **rozdílné množiny nalezených dokumentů** při shodně zadaných kritérií.
- Přidávání nové funkcionality a úpravy ve stávajících aplikacích je náročnější včetně většího rizika vzniku chyb.

3 Požadavky společnosti LMC na nový search engine

Vzhledem k neustálému vývoji nových služeb, narůstajícímu počtu prohledávaných dat a plánovanému rozvoji společnosti LMC lze rozdělit požadavky na nové vyhledávání do následujících kategorií:

3.1 Podporované technologie v LMC

- Pokud bude vybrané řešení provozováno a spravováno v LMC, je nutné, aby splňovalo požadavky na technologie uvedené v kapitole *Architektura a implementace*.

3.2 Využitelnost v rámci rozdílných aplikací

- Vzhledem k rozsáhlému množství provozovaných služeb popsaných v kapitole *Využití search enginu v LMC* je nutné, aby bylo možné jednoduše

zprovoznit nové instance search enginu s odlišným obsahem, nastavením a strukturou dat.

- Možnost zakomponovat nové řešení search enginu i do automatizovaného vyhledávání, které je popsáno v kapitole *Aplikace Agenti*.

3.3 Výkonnost a škálovatelnost

- Průměrná doba zpracování požadavku na vyhledávání je < 1s.
- Lze provozovat více instancí search enginu pro zvyšování výkonu.

3.4 Management a správa aplikace

- Stav aplikace včetně možnosti napojení na monitoring.
- Snadné přidávání nových strukturovaných kritérií a nastavování relevancí.
- Možnost vyhledávat pouze nad vybranou množinou dokumentů.
- Možnost sledování a analýzy dat.

3.5 Požadavky na vyhledávání

- Fulltextové vyhledávání podporuje vyhodnocování relevancí s možnostmi odlišit důležitost v různých částech dokumentu.
- Podpora různých jazyků, které jsou již dnes v aplikacích provozovaných v LMC využívány. Mezi ty patří především čeština, slovenština, němčina, angličtina a následně dalších 8 evropských jazyků. Podporou je zde míněno nejen jejich správné zobrazování, ale také možností práce s daným jazykem (Morfologie).
    - Stemming – jedná se o rozklad slova s cílem nalezení jeho kořenu pomocí různých algoritmů.
    - Možnosti napojení různých slovníků včetně využití přiřazování synonym.
- Možnost vytvářet nové kategorie a jejich provázanosti např. mezi Obory a profesemi. Mezi jednotlivými kategoriemi a jejich položkami je možné stanovovat váhy a tím ovlivňovat výslednou relevanci kritérií.
- Možnost využití odlišných algoritmů vyhodnocení pro strukturovaná kritéria oproti standardně využívaným ve fulltextovém vyhledávání.

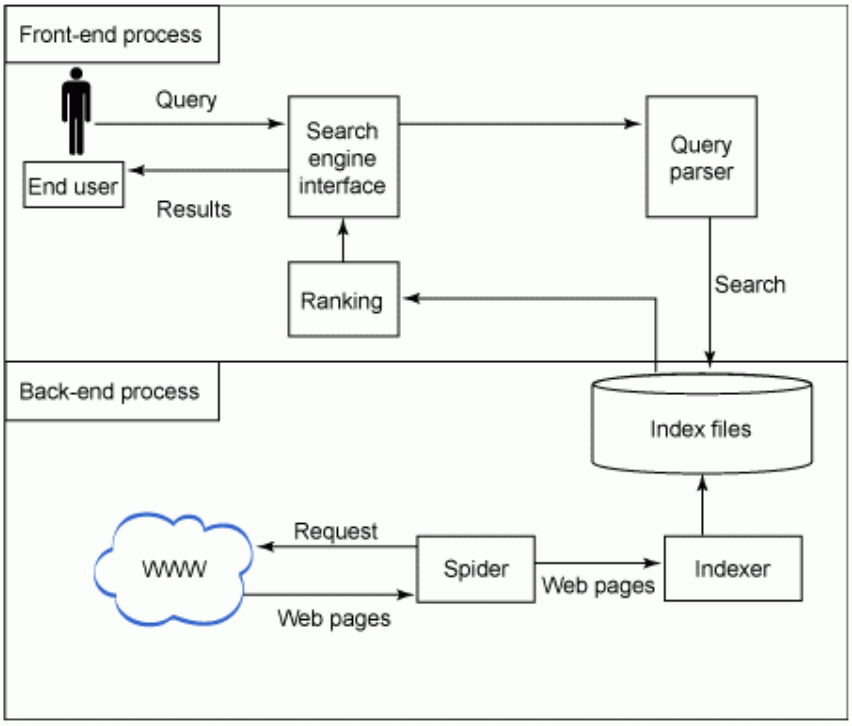
- Podpora zvýraznění nalezených klíčových slov ve výpisu.
- Zohlednění různých obchodních modelů podle aplikace, pro kterou je search engine provozovaný. Např. na serveru jobs.cz není na výpisu umožněno využít různé předplacené zvýhodnění dokumentů a vše je tedy ponecháno na přesnosti zadání dokumentu, požadovaných kritérií a následném vyhodnocení relevancí. Narozdíl od portálu prace.cz, kde je vše postaveno na striktním porovnání, prioritách předplacených zvýraznění a řazení dle data.

3.6 Cena a podpora dodavatele

- Pomoc s osvojením si nového řešení ze strany dodavatele.
- Možnost ovlivnit další vývoj aplikace.
- Cena řešení, kde je zohledněna jak počáteční investice tak investice v dalších letech.

4 Koncepce search enginu

V této kapitole představím principy fungování web search enginu. Jedním z nich je i open source projekt Apache Lucene (lucene.apache.org), jehož dokumentace je také jedním ze zdrojů, které zde byly využity. Architekturu search enginu můžeme rozdělit dle procesů na dvě části front-end a back-end, jak je vidět na obrázku 4.



Obrázek 4 - Architektura search enginu (zdroj [Zhou_06 www])

4.1 Back-end procesy

V části back-end dochází nejprve k volání robota zvaného **Crawler** (někdy také Spider), který podle přesně nastavených pravidel stahuje obsah jednotlivých stránek, provádí jejich komprimaci a ukládá je do repository. V případě, kdy je doba potřebná pro stažení a tedy i aktualizaci určeného obsahu příliš dlouhá, je vhodné využít více instancí Crawleru.

Nad repository je spuštěn HTML analyzátor, který uložené soubory dekomprimuje, rozparsuje podle HTML tagů a uloží jejich obsah v textové podobě do DB nebo ve

formě souborů na disk. V případě podnikových aplikací není vždy nutné Crawler využít, protože data již mohou být ve vhodném formátu.

Před samotným procesem vytváření indexu prochází vstupní data **lexikální analýzou** a filtrováním, jejichž cílem je extrahovat informace vhodné pro indexování a vytvořit z nich tzv. tokeny. Každý token reprezentuje slovo v textu a obsahuje kromě samotné hodnoty slova, také další metadata, např. počáteční a koncovou polohu v textu nebo datový typ.

Součástí Lucene jsou následující analyzátory:

- SimpleAnalyzer - odděluje od textu znaky, které nejsou v abecedě a převádí znaky na malá písmena.
- StopAnalyzer - odděluje od textu znaky, které nejsou v abecedě, převádí znaky na malá písmena a vyřadí slova uvedená v seznamu zakázaných slov.
- StandardAnalyzer - vytváří tokeny ze vstupního textu.
- WhitespaceAnalyzer - eliminuje prázdná místa v textu.

Po analýze textu je v průběhu **indexování**, každý token uložen do indexu jako term. Term reprezentovaný textem je společně s identifikátorem dokumentu základním prvkem indexu. Analyzovaná data jsou následně ukládána do struktury zvané **invertovaný index**, který již umožňuje rychlé vyhledávání. Soubor s invertovaným indexem obsahuje pro každý term seznam odkazů na všechny dokumenty, ve kterých se vyskytuje.

	Dok1	Dok2	Dok3
Zaměstnanec	0	0	1
Zaměstnavatel	0	1	0
Agentura	1	1	0

→

	Dok1	Dok2	Dok3
Zaměstnanec	0	0	3
Zaměstnavatel	0	2	0
Agentura	1	2	0

Obrázek 4.1 – Ukázka tvorby invertovaného indexu ze vstupní matice

Schématicky lze strukturu Lucene indexu rozdělit na následující části:

Index > Segmenty > Dokumenty > Pole

Index se skládá z nezávislých segmentů, z nichž každý obsahuje stanovený počet dokumentů. Dokument se skládá ze sekvence polí, které představují data dokumentu. Každý nový dokument je potom reprezentován novým indexovým segmentem, který je následně v rámci optimalizace rychlosti vyhledávání slučován do segmentů větších.

4.2 Front-end procesy

V části front-end zadává uživatel dotaz pomocí vstupního rozhranní. Následuje **syntaktická analýza**, při které dojde k převedení dotazu uživatele na výraz. Výraz je rozdělen na operátory a termy, které mohou být dvojího druhu:

- Jednoslovné termy (např. “Java“)
- Víceslovné termy neboli fráze vkládané do uvozovek (např. “Java developer“)

Termy mohou být pro složitější dotazy kombinovány s operátory (např. “Java programmer“ AND (“Praha“ OR “Brno“))

Lucene podporuje různé **syntaxe zadávání dotazů**:

- Field - hledání v určitém poli (např. Profese: „Java developer“)
- Wildcards - nahrazování znaků pomocí “*“ pro skupinu znaků nebo “?“ pro jeden znak (např. Prah* ; Ja?a)
- Fuzzy search - hledání podobných slov (např. test~)
- Proximity search - hledá zadané termy vzdálené od sebe v maximálně určeném rozsahu (např. “jakarta apache“~10)
- Range search - vyhledává ve stanoveném rozsahu (např. plat:[20000 TO 30000])
- Boosting a Term - ovlivňuje relevanci termu (např. Java^4 Programmer). Relevantnost se zvyšuje pokud je číslo za “^” > 1 a snižuje v rozsahu od 1 do 0.
- Boolean operátory - AND, OR, NOT, “+”,”-” (např. +jakarta OR apache AND website)
- Seskupování pomocí závorek (např: title:(+Java +“Praha Prosek“))

Jakmile byl dotaz analyzován a rozložen, dojde k **vyhledávání jednotlivých termů v indexu**. Výsledkem vyhledávání je množina záznamů v XML formátu které reprezentují dokumenty obsahující hledané termy. Každý záznam se skládá z identifikátoru dokumentu, vypočtené relevance a případně dalších metadat. Podle údajů v seznamu lze již snadno dohledat odpovídající dokumenty, vybrat z nich případné další údaje, které nebyly indexovány a přetransformovat výstup do podoby, kterou je možné zaslat zpět uživateli.

5 Možnosti řešení search enginu

V rámci ČR je na trhu jen několik firem, které se zabývají **prodejem** nebo **kompletním outsourcingem** svého vlastního řešení search enginů. Jedná se především o společnost Jyxo s.r.o. a Netcentrum s.r.o. Ve výběru byly proto zohledněny i některé zahraniční společnosti - Actonomy NV a SearchBlox Software, Inc. Ve světě existuje samozřejmě mnoho dalších dodavatelů např. FAST, ale vzhledem k nedostatku dostupných informací nebyly do výběru zahrnuty.

Mezi další varianty patří realizace search enginu pomocí **open source** technologií, kde je možné si aplikaci upravit na vlastní náklady a podle svých požadavků bez dalších poplatků spojených s licenční politikou.

5.1 Varianty se zakoupením hotového řešení

Zakoupení již hotového řešení představuje pro LMC nákup aplikace search enginu, podle licenční politiky dodavatele, která bude provozována na vlastním HW. V rámci této varianty mohou být s dodavatelem smluveny dodatečné služby mezi které patří např. administrace, aktualizace databází, upgrade software, dohled apod.

Společnosti nabízející vlastní řešení:

- **Varianta 1 (V1)** Jyxo s.r.o. (dále jen Jyxo) – (www.jyxo.cz), která provozuje vlastní fulltextový search engine shodného názvu.

- **Varianta 2 (V2)** Actonomy – (www.actonomy.com) je zahraniční společnost nabízející vlastní řešení s názvem xMP, které kombinuje vyhledávání přes strukturovaná data s fulltextovým search enginem.
- **Varianta 3 (V3)** SearchBlox Software, Inc. (dále jen SearchBlox) – (www.searchblox.com) je zahraniční společnost nabízející vlastní řešení postavené na technologii Apache Lucene.

5.2 Varianta formou pronájmu

Forma pronájmu představuje z pohledu LMC pouze zakoupení licence, která umožňuje využívat API daného search enginu provozovaného na HW dodavatele. Zodpovědností dodavatele jsou potom veškeré činnosti spojené s provozem aplikace.

- **Varianta 4 (V4)** Jyxo s.r.o. (dále jen Jyxo) - (www.jyxo.cz), která provozuje vlastní fulltextový search engine shodného názvu.

5.3 Varianty využívající open source technologie

Výběr některé open source technologie by pro LMC znamenal především náklady na přepis současného řešení s tím, že aplikace by byla provozována na vlastním HW a taktéž spravována v LMC. U vybraných open source projektů je třeba si uvědomit, že se jedná především o fulltextové vyhledávače, které neobsahují veškerou požadovanou funkčnost, ale jsou otevřeny pro další úpravy. Úpravy však mohou být značně náročné a proto je výhodnější využít firem, které již tyto technologie využívají a mají potřebné zkušenosti s jejich implementací. Jedná se např. o společnost Conlegere s.r.o. nebo Kyberie s.r.o.

Přehled variant využívajících open source řešení:

- **Varianta 5 (V5)** Apache Lucene – (lucene.apache.org) je výkonný textový search engine dostupný ve formě knihovny napsaný v programovacím jazyce Java. Implementováno společností **Conlegere s.r.o.** (dále je Conlegere)
- **Varianta 6 (V6)** Apache Lucene – (lucene.apache.org) představuje řešení postavené na shodné technologii jako varianta 5, ale implementace je ponechána na společnosti **Kyberie s.r.o.** (dále jen Kyberie)

6 Stanovení metrik

Tato kapitola je úvodem do problematiky metrik, jejich členění a následné stanovení metrik vlastních, které budou využity v kapitole *Porovnání a návrh vhodného řešení* při rozhodování mezi jednotlivými variantami realizace search enginu.

Podle [Ucen_01] *je metrika „přesně vymezený finanční či nefinanční ukazatel či hodnotící kritérium, které je používáno k hodnocení úrovně efektivnosti konkrétní oblasti řízení podnikového výkonu a jeho efektivní podpory prostředky IS/ICT. Skupinu metrik sdružených za určitým cílem (tzn. vztahujících se ke konkrétní oblasti, procesu či projektu) nazýváme „portfolio metrik“.*

Pro vybrané metriky je nutné stanovit:

- hodnoty, které budou předmětem zkoumání,
- veličiny, ve kterých budou hodnoty uvedeny,
- kroky vedoucí k vyhodnocení výsledků.

6.1 Rozdělení metrik

Obecně lze metriky rozdělit na dvě skupiny - tvrdé a měkké. Tvrdé metriky jsou měřitelné ukazatele, které nevyžadují téměř žádné dodatečné náklady a často se dají převést na finanční ukazatele. Mezi tvrdé metriky lze mimo ukazatelů považovat také indikátory, které si lze představit jako určitou horní nebo spodní mez. Jako příklad tvrdých metrik lze uvést celkové náklady na implementaci služby, maximální dobu odezvy aplikace nebo procentuální stanovení dostupnosti služby.

Měkké metriky představují kvalitativní ukazatele, u kterých již není měření tak snadné a často je spojeno s dodatečnými náklady. Obvykle je jejich zjišťování prováděno v rámci různých průzkumů s využitím dotazníků nebo anket. Příkladem měkké metriky může být spokojenost zákazníků s využívanými službami nebo hodnocení kvality školení apod. Měkké metriky lze převést na metriky tvrdé, a to postupem přiřazování číselných hodnot stupňům míry. S číselnými hodnotami lze potom snadno provádět matematické operace.

Podle [Ucen_01] mají metriky několik atributů, z nichž některé jsou využity v rámci stanovení vlastních metrik. Mezi ně patří především identifikace metriky, typ a název

metriky, definice metriky, účel metriky, vzorec a výpočet dat, měrná jednotka a interpretace naměřených dat. Skupině metrik, které se vztahují ke konkrétní oblasti, procesu či projektu, říkáme "portfolio metrik".

6.2 Použité metriky

Pro porovnání jednotlivých variant definovaných v kapitole *Možnosti řešení search enginu*, byly zvoleny následující metriky:

- metrika náročnosti implementace a správy search enginu,
- metrika rozsahu požadované funkčnosti,
- metrika výkonnosti search enginu,
- metrika celkových nákladů.

6.2.1 Metrika náročnosti implementace a správy search enginu

Pomocí této metriky lze posoudit náročnost dílčích činností a odvodit dobu implementace v rámci každé z variant realizace search enginu. Mezi tyto činnosti patří **analýza, programování, instalace, testování a správa aplikace,** přičemž každá má stanoven maximální podíl na jejich celkovém součtu (viz vzorec). Náročnost jednotlivých činností byla odvozena z předchozích zkušeností na obdobných projektech.

Název	Metrika náročnosti implementace a správy search enginu
Definice	Metrika pro posouzení celkové náročnosti činností, jimiž jsou: analýza, programování, instalace, testování a správa aplikace, v rámci každé z variant realizace search enginu.
Měřená jednotka	Procenta
Vzorec	$CNI = \sum_{i=1}^{n} Ci \cdot k$ Kde CNI = celková náročnost implementace % C[i] = výčet činností, n=počet činností C[1] = (20 %) = náročnost na analýzu (v %) C[2] = (40 %) = náročnost na programování (v %) C[3] = (5 %) = náročnost na instalaci (v %) C[4] = (10 %) = náročnost na konfiguraci (v %) C[5] = (20 %) = náročnost na testování (v %) C[6] = (5 %) = náročnost na správu aplikace (v %)

	k = koeficient nabývající hodnoty od 0 do 1 v závislosti na náročnosti dané činnosti z pohledu firmy LMC Výsledná hodnota CNI tedy může v rámci každé varianty nabývat hodnot od 0 % do 100 %.
Interpretace hodnot	Čím nižší je celkový součet CNI, tím je nižší celková náročnost a předpokládaná doba implementace aplikace v rámci každé varianty.

Tabulka 6.2.1 - definice metriky náročnosti implementace a správy

6.2.2 Metrika rozsahu požadované funkčnosti

Pomocí této metriky lze posoudit rozsah poskytování požadované funkčnosti search enginu, který ovlivňuje možnosti správného vyhodnocení dotazu v rámci každé z variant. Mezi tyto požadavky patří **relevantní fulltext, podpora českého jazyka, podpora jiných jazykových mutací, zvýraznění nalezených klíčových slov, kategorizace s vazbami mezi kritérii a využití rozdílných algoritmů pro strukturovaná kritéria,** přičemž každá má stanoven podíl na jejich celkovém součtu (viz vzorec). Podíly jsou stanoveny na základě důležitosti jednotlivých funkčností z pohledu LMC,

Název	Metrika rozsahu požadované funkčnosti
Definice	Metrika pro posouzení rozsahu poskytované funkčnosti, jíž jsou: relevantní fulltext, podpora českého jazyka, podpora jiných jazykových mutací, zvýraznění nalezených klíčových slov, kategorizace a provázanost kritérií, využití rozdílných algoritmů pro strukturovaná kritéria v rámci každé z variant realizace search enginu.
Měřená jednotka	Procenta
Vzorec	$CF = \sum_{i=1}^{n} Fi \cdot k$ Kde CF = celková požadovaná funkčnost (v %) F[i] = výčet požadovaných funkčnosti, n=počet funkčností F[1] = (40 %) = podpora relevantního fulltextu (v %) F[2] = (20 %) = podpora českého jazyka (v %) F[3] = (10 %) = podpora jiných jazyků (v %) F[4] = (5 %) = podpora zvýraznění klíčových slov (v %) F[5] = (15 %) = podpora kategorizace a provázanosti kritérií (v %) F[6] = (10 %) = využití rozdílných algoritmů pro strukturovaná kritéria (v %) k = koeficient nabývající hodnoty od 0 do 1 v závislosti na tom, do

	jaké míry je daná funkčnost podporována. Výsledná hodnota CF tedy může v rámci každé varianty nabývat hodnot od 0 % do 100 %.
Interpretace hodnot	Čím vyšší je celkový součet CF, tím se zvyšují možnosti search enginu nalézt odpovídající množinu výsledků k zadanému dotazu.

Tabulka 6.2.2 - definice metriky rozsahu požadované funkčnosti

6.2.3 Metrika výkonnosti search enginu

Pomocí této metriky lze posoudit výkonnost variant realizace search enginu. Metrika je specifikována jako **průměrná doba obsloužení požadavku o N kritériích, při počtu M paralelních dotazů nad množinou X dokumentů**. Zjištění průměrné doby je podrobněji popsáno v kapitole *Výkonnost search enginů*.

Název	Metrika výkonnosti search enginu
Definice	Metrika je specifikována jako průměrná doba obsloužení požadavku o N kritériích, při počtu M paralelních dotazů nad množinou X dokumentů.
Měřená jednotka	Čas v milisekundách
Vzorec	Zjištění průměrné doby je podrobněji popsáno v kapitole *Výkonnost search enginů*.
Interpretace hodnot	Čím nižší je naměřená průměrná doba obsloužení požadavku tím je search engine výkonnější.

Tabulka 6.2.3 - definice metriky výkonnosti search enginu

6.2.4 Metrika celkových nákladů

Pomocí této metriky lze posoudit celkové náklady na implementaci každé z variant realizace search enginu. Mezi jednotlivé náklady jsou zahrnuty tyto položky: **náklady na analýzu**, **náklady na programování**, **náklady na instalaci a konfiguraci**, **náklady na testování**, **náklady na správu aplikace**, **cena řešení** dle typu licenční politiky. Podíl nákladů na jednotlivých činnostech je odvozen z metriky náročnosti implementace a správy search enginu, a je doplněn o cenové nabídky oslovených dodavatelů. Vzhledem k tomu, že LMC již vlastní vhodné HW vybavení nebyly náklady na HW zařazeny do celkového výpočtu.

Název	Metrika celkových nákladů
Definice	Výpočet celkových nákladů, do kterých jsou zahrnuty tyto položky: náklady na analýzu, náklady na programování, náklady na instalaci a konfiguraci, náklady na testování, náklady na správu aplikace, cena řešení dle typu licenční politiky v rámci každé varianty.
Měřená jednotka	Cena v Kč
Vzorec	$CN = \sum_{i=1}^{n} Ni$ Kde CN = celkové náklady N[i] = výčet jednotlivých nákladů, n=počet zohledněných nákladů N[1] = náklady na analýzu N[2] = náklady na programování N[3] = náklady na instalaci N[4] = náklady na konfiguraci N[5] = náklady na testování N[6] = náklady na správu N[7] = náklady na nákup licencí
Interpretace hodnot	Čím nižší je hodnota celkových nákladu CN, tím je varianta výhodnější.

Tabulka 6.2.4 - definice metriky celkových nákladů

7 Výkonnost search enginů

Tato kapitola poskytuje základní informace týkající se problematiky porovnání výkonnosti různých variant search enginů. Samotné porovnání jednotlivých variant pomocí Metriky výkonnosti search enginů nebylo z důvodů náročnosti zatím provedeno, a proto zde budou alespoň představeny výsledky měření, které bylo možné k vybraným řešením od dodavatelů získat.

7.1 Požadavky a postup testování

Přípravu na test a vlastní testování lze rozdělit do několika částí.

- **HW vybavení,** na kterém budou testy prováděny. Doporučením je samostatný server, na kterém nejsou v době provádění testu spuštěny žádné další náročnější aplikace. V případě, že se jedná o server, který je součástí produkčního prostředí, potom, je-li to možné, by měl v době testu být z produkčního prostředí vyřazen a nebo by měl test probíhat alespoň v době,

kdy je prostředí co nejméně vytěžováno.

Doporučená konfigurace serveru vhodného pro zátěžový test variant uvedených v této práci je např. INTEL Quad-Core XEON 5310, 1.6 GHz, 4GB RAM a diskem SAS s 15000 rpm.

- **SW vybavení** by mělo odpovídat především již vyzkoušeným řešením doporučeným dodavatelem, a to včetně postupu pro instalaci search enginu. Vzhledem k tomu, že všechny varianty, které by byly provozovány na straně LMC jsou dostupné v jazyce Java, je vhodné využít software blíže specifikovaný v kapitole *Architektura a implementace* v části preferované technologie s tím, že pro samotné měření lze využít např. aplikaci Apache JMeter, která je volně dostupná na internetu (http://jakarta.apache.org).
- **Stanovení cíle měření a nastavení testu** může být odlišné v závislosti na definici použité metriky. Obecně lze říci, že pro testování search enginů potřebujeme nastavit aplikaci JMeter pro generování http dotazů a kontroly rychlosti jejich odpovědí ve stanoveném čase. Výstupy testování lze zobrazit v přehledných reportech včetně možnosti vygenerování grafů. Http dotaz představuje odeslání všech zadaných kritérií z vyhledávacího formuláře metodou POST nebo GET na server.

  Test by měl probíhat nad předem známým počtem dokumentů v indexu např. 10.000, 100.000 a 1.5 mil. s tím, že jednotlivé dokumenty se od sebe liší. Dostatečnou množinu dokumentů lze vygenerovat např. pomocí náhodného generování řetězců. Dalšími vstupními parametry, které jsou podstatné pro test jsou:

  - Počet zadaných kritérií, např. 1, 10 a 20, přičemž hodnota 10 splňuje většinu současných dotazů při vyhledávání na internetu.
  - Počet dokumentů zobrazených na výpise, např. 40 a 100.
  - Počet paralelních uživatelů 1, 10 a 100 bez nutnosti definovat prodlevu mezi jednotlivými dotazy.
  - Doba, po kterou bude test spuštěn např. 10 minut.

7.2 Výkonnost jednotlivých variant

Jak již bylo zmíněno v předešlé kapitole, test podle všech požadavků potřebných pro porovnání se zatím nepodařilo realizovat, a proto jsou níže prezentována alespoň některá čísla uváděná samotnými dodavateli řešení.

7.2.1 Výkonnost varianty V5 – Conlegere s.r.o.

Použitý HW: Intel(R) Xeon(R), CPU E5310 @ 1.60GHz, 4GB RAM, Disk SAS 15.000 rpm.

Specifikace testu: dokumentů v indexu 100.000, dokumentů na výpisu 40, paralelních uživatelů 10 a 100, kritérií v dotazu 1 a 25, přičemž bylo eliminováno využití cache použitím generování náhodných dotazů.

Výstupy: průměrná odezva při 10 paralelních uživatelích byla pro 1 kritérium 65 ms a pro 25 kritérií 94 ms. Pokud byl zvýšen počet uživatelů na 100 stále byla odezva do 1000 ms, což je jednoznačně nejlepší výsledek ve všech dostupných testech.

7.2.2 Výkonnost varianty V6 – Kyberie s.r.o.

Použitý HW: server SunFire V40z, 4xCPU Opteron 850 2.4 GHz, 8GB RAM, disk 2x 136GB SCSI RAID 1.

Specifikace testu: dokumentů v indexu 10.000, paralelních uživatelů 10 a 100, kritérií v dotazu 1 a 10, přičemž bylo eliminováno využití cache použitím generování náhodných dotazů.

Výstupy: průměrná odezva při 10 paralelních uživatelích byla pro 1 kritérium 516 ms a pro 10 kritérií 88 ms. Pokud byl zvýšen počet uživatelů na 100 zvedla se průměrná odezva až nad 5000 ms.

7.2.3 Výkonnost ostatních variant

- Společnost Jyxo pro variantu V1 a V4 uvádí průměrnou dobu zpracování dotazu nad množinou 200.000 dokumentů 50ms.

- Společnost Actonomy u varianty V2 uvádí maximální dobu zpracování dotazu nad množinou 150.000 dokumentů při 12 zadaných kritériích do 1000 ms.
- Společnost SearchBlox Software, Inc. pro variantu V3 neposkytuje přesné údaje, ale vzhledem k tomu, že využívá technologii Apache Lucene, kterou zde již hodnocenou máme, lze soudit, že rychlost je především otázkou správné konfigurace a použitého HW vybavení.

8 Porovnání a návrh vhodného řešení

Obsahem poslední kapitoly je aplikace jednotlivých metrik definovaných v kapitole *Použité metriky* na jednotlivé varianty popsané v kapitole *Možnosti řešení search enginu* a návrh optimálního řešení pro LMC. Při aplikaci metrik je vycházeno z odhadů, které nejsou podloženy celkovou technickou analýzou, a proto je lze považovat spíše jako přibližný rámec, který byl sestaven na základě zkušeností z projektů obdobného rozsahu.

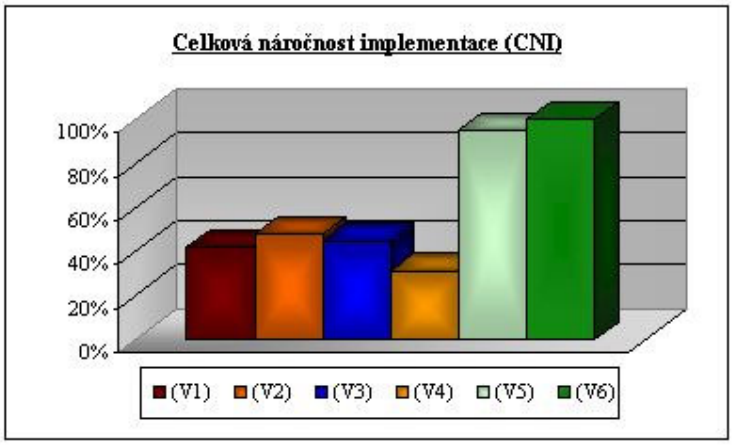
8.1 Aplikace metriky náročnosti implementace a správy search enginu

Na základě metriky náročnosti implementace a správy search enginu definované v kapitole *Použité metriky* lze nyní vyjádřit celkový procentuální součet náročnosti jednotlivých činností pro všechny varianty uvedené v kapitole *Možnosti řešení search enginu*. Výstup je prezentován v následující tabulce, přičemž procentuální vyjádření uvedené u názvu jednotlivých činnosti představují jejich odhadovanou maximální náročnost a procenta uvedená u jednotlivých variant potom jejich odhadovanou reálnou hodnotu.

	(V1)	(V2)	(V3)	(V4)	(V5)	(V6)
C[1] = (20%) = náročnost na analýzu	8	10	8	5	15	20
C[2] = (50%) = náročnost na programování	10	10	10	10	50	50
C[3] = (3%) = náročnost na instalaci	5	5	5	0	3	3
C[4] = (3%) = náročnost na konfiguraci	2	3	2	1	3	3
C[5] = (15%) = náročnost na testování	15	15	15	15	15	15
C[6] = (10%) = náročnost na správu aplikace	2	5	5	0	9	9
CNI = celková náročnost implementace (v %)	**42**	**48**	**45**	**31**	**95**	**100**

Tabulka 8.1 – aplikace metriky náročnosti implementace a správy search enginu

Pro přehlednost jsou převedeny celkové součty týkající se náročnosti implementace jednotlivých variant do grafické podoby.



Graf 8.1 – Výsledek metriky náročnosti implementace a správy search enginu

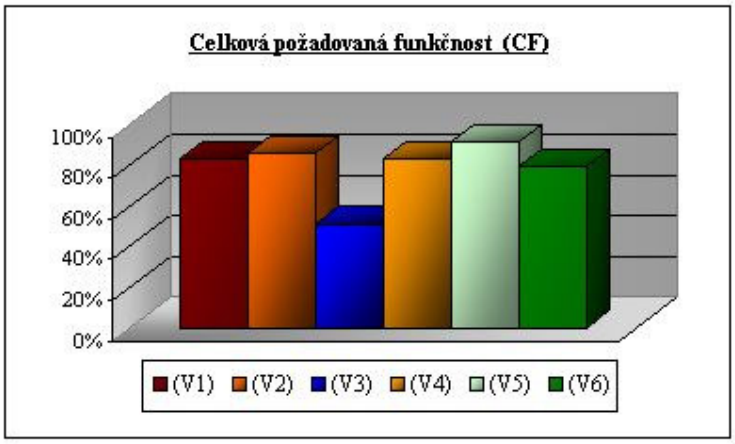
8.2 Aplikace metriky rozsahu požadované funkčnosti

Na základě metriky rozsahu požadované funkčnosti definované v kapitole *Použité metriky* lze nyní vyjádřit celkový procentuální součet míry vyžadované funkčnosti pro všechny varianty uvedené v kapitole *Možnosti řešení search enginu*. Výstup je prezentován v následující tabulce, přičemž procentuální vyjádření uvedené u názvu jednotlivých požadavků na funkcionalitu představuje jejich vzájemnou váhu a maximální ohodnocení. Procenta uvedená u jednotlivých variant jsou jejich odhadovanou reálnou hodnotou.

	(V1)	(V2)	(V3)	(V4)	(V5)	(V6)
F[1] = (40%) = podpora relevantního fulltextu	40	40	40	40	40	40
F[2] = (20%) = podpora českého jazyka	20	10	0	20	20	10
F[3] = (10%) = podpora jiných jazyků	6	8	7	6	8	7
F[4] = (5%) = podpora zvýraznění klíčových slov	5	5	4	5	5	4
F[5] = (15%) = podpora kategorizace a provázanosti kritérií	10	15	0	10	8	8
F[6] = (10%) = využití rozdílných algoritmů pro strukturovaná kritéria	2	8	0	2	10	10
CF = celková požadovaná funkčnost (v %)	**83**	**86**	**51**	**83**	**91**	**79**

Tabulka 8.2 – aplikace metriky rozsahu požadované funkčnosti

Pro přehlednost jsou převedeny celkové součty týkající se rozsahu požadované funkčnosti jednotlivých variant do grafické podoby.



Graf 8.2 – Výsledek metriky rozsahu požadované funkčnosti

8.3 Aplikace metriky výkonnosti search enginu

Metrika výkonnosti search enginu popsaná v kapitole *Použité metriky* nebyla z důvodu náročnosti na přípravu, která je blíže specifikována v kapitole *Požadavky a postup testování* aplikována. Pro posouzení rychlosti byly v kapitole *Výkonnost jednotlivých variant* uvedeny testy poskytnuté dodavateli řešení.

8.4 Aplikace metriky celkových nákladů

Na základě metriky celkových nákladů definované v kapitole *Použité metriky* lze nyní vyjádřit celkový součet nákladů pro všechny varianty uvedené v kapitole *Možnosti řešení search enginu*. Nejprve byly stanoveny hodinové sazby a odhady pracnosti pro variantu V6, která je podle metriky náročnosti implementace a správy search enginu maximálně náročná na všechny uvedené činnosti a znamená tedy i nejvyšší náklady na tyto činnosti z pohledu LMC (tabulka 8.4a).

	Kč/hod.	**Dny**	**Náklady**	%
N[1] = náklady na analýzu	1500	23	276.000	0,20
N[2] = náklady na programování	1500	58	696.000	0,50
N[3] = náklady na instalaci	1500	3	36.000	0,03
N[4] = náklady na konfiguraci	1500	3	36.000	0,03
N[5] = náklady na testování	1000	17	136.000	0,15
N[6] = náklady na správu	1500	12	144.000	0,10
Celkové náklady	-	**116**	**1.324.000**	**100%**

Tabulka 8.4a – odhad maximální pracnosti
a nákladů k jednotlivým činnostem

Suma nákladů v tabulce 8.4a představuje částku, která je následně zohledněna při výpočtu dílčích nákladů v tabulce 8.4b podle míry náročnosti stanovené v metrice náročnosti implementace a správy search enginu.

Náklady spojené s licenční politikou jsou nejprve spočítány pouze na jeden rok při předpokladu, že aplikace bude provozována na pěti serverech s průměrným počtem dotazů 6 milionů za měsíc. Tyto čísla vycházejí ze statistik současného řešení.

- V1 - měsíční poplatek činí 16.000 Kč za jeden server a další 2.000 Kč za každý další. Roční náklady lze tedy vyjádřit vztahem:
  NV1= (16.000 + (4 x 2.000) x 12) = 288.000 Kč
- V2 – roční poplatek činí 13.500 EUR, což lze po přepočtu vyjádřit vztahem:
  NV2 = (13.500 x 28) = 378.000 Kč
- V3 – roční poplatek činí 7.499 USD, což lze po přepočtu vyjádřit vztahem:
  NV3 = (7.499 x 18) = 134.982 Kč
- V4 - měsíční poplatek činí 15.000 Kč za 500.000 dotazů a následně 12 Kč za každých dalších 1.000 dotazů. Roční náklady lze tedy vyjádřit vztahem:
  NV4= (15.000 + (12 x 5.500)) x 12 = 972.000 Kč
- Varianta V5 a V6 není licenčně zpoplatněna

Vzhledem k značným rozdílům na náklady v počátečním roce a následně v dalších letech byl do tabulky přidán i výpočet nákladů na dobu 3 let, který je tvořen náklady na první rok a v následujících letech, již pouze platbami za správu aplikace a cenu licencí.

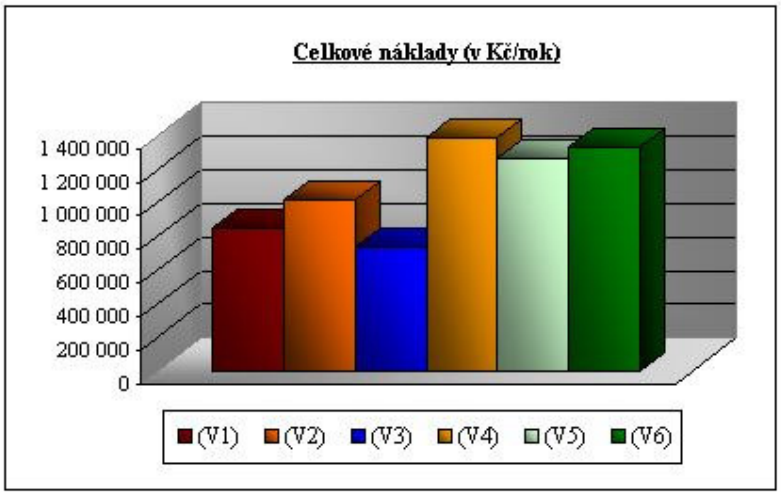
Náklady za 3 roky lze vyjádřit vztahem: CN3 = CN + ((N[6]+[N7]) x 2)

Aplikace metriky celkových nákladů je prezentována v následující tabulce.

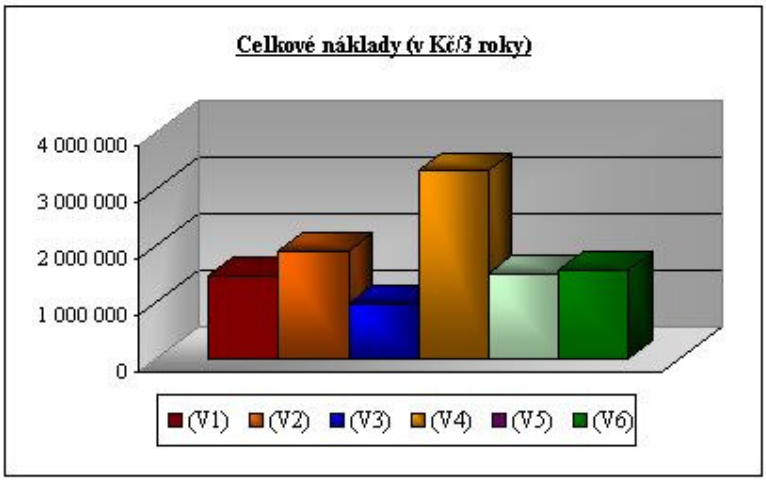
	(V1)	(V2)	(V3)	(V4)	(V5)	(V6)
N[1] = náklady na analýzu	105.920	132.400	105.920	66.200	198.600	264.800
N[2] = náklady na programování	132.400	132.400	132.400	132.400	662.000	662.000
N[3] = náklady na instalaci	66.200	66.200	66.200	0	39.720	39.720
N[4] = náklady na konfiguraci	26.480	39.720	26.480	13.240	39.720	39.720
N[5] = náklady na testování	198.600	198.600	198.600	198.600	198.600	198.600
N[6] = náklady na správu	26.480	66.200	66.200	0	119.160	119.160
N[7] = náklady na nákup licencí	288.000	378.000	134.982	972.000	0	0
CN = náklady (v Kč / rok)	**844.080**	**1.013.520**	**730.782**	**1.382.440**	**1.257.800**	**1.324.000**
CN3 = náklady (v Kč / 3 roky)	**1.473.040**	**1.901.920**	**957.669**	**3.326.440**	**1.496.120**	**1.562.320**

Tabulka 8.4b – aplikace metriky celkových nákladů

Pro přehlednost byly převedeny celkové náklady na jednotlivé varianty v období 1 a 3 roky do grafické podoby.



Graf 8.4b – Výsledek metriky celkových nákladů na jeden rok



Graf 8.4c – Výsledek metriky celkových nákladů na 3 roky

8.5 Závěry

Závěrečná kapitola je souhrnem všech poznatků vycházejících z použitých metrik s účelem doporučit vhodnou variantu search enginu pro společnost LMC. Při výběru byly zohledněny i požadavky na nové řešení stanovené v kapitole *Požadavky LMC na nový search engine*.

Pro snadnější porovnání výsledků byla vytvořena tabulka 8.5, ve které jsou pomocí bodového systému ohodnoceny jednotlivé varianty v rámci každé z použitých metrik. Body jsou stanoveny v rozmezí 1-5 a odpovídají pořadí výsledků zjištěných při aplikaci metrik. Čím vyšší je celkový součet bodů u dané varianty, tím je varianta považována za vhodnější.

	(V1)	(V2)	(V3)	(V4)	(V5)	(V6)
CNI = celková náročnost implementace	5	3	4	6	2	1
CF = celková požadovaná funkčnost	3	5	1	3	6	2
CN = náklady (v Kč / rok)	5	4	6	1	3	2
CN3 = náklady (v Kč / 3 roky)	5	2	6	1	4	3
Celkový součet	**18**	**14**	**17**	**11**	**15**	**8**

Tabulka 8.5 – Porovnání výsledků metrik

Na základě předchozích výpočtu lze konstatovat následující závěry:

- Jak je na první pohled patrné varianta (V6), která využívá open source technologii Apache Lucene implementovanou společnosti Kyberie je nejvíce náročná na její implementaci, což souvisí také s vyššími počátečními náklady a neposkytuje dostatečný rozsah požadované funkčnosti.

  Pokud navíc přihlédneme k výsledkům výkonnostních testů uvedených v kapitole *Výkonnost jednotlivých variant*, tak je ve srovnání s implementací shodné technologie společností Conlegere výrazně horší, a to především při zvyšujícím se počtu paralelních dotazů. Je to způsobeno především proto, že společnost Kyberie nemá s touto technologií zatím tolik zkušeností jako její uvedený konkurent, což se v budoucnu může změnit. Vzhledem k uvedeným skutečnostem **nelze variantu (V6) implementovanou společností Kyberie doporučit**.
- Varianta (V4), která je jako jediná realizována kompletním outsourcingem search enginu nabízeného společností Jyxo, je ideální z pohledu náročnosti implementace a v případě kdy není možné využít vlastní HW vybavení, což ovšem není případ společnosti LMC.

  Tato varianta neposkytuje veškerou požadovanou funkčnost v dostatečném rozsahu, a proto by vyžadovala dodatečné náklady na provedení úprav. V požadované funkčnosti vyniká především při práci s českým jazykem. Největší nevýhodou tohoto řešení je cena, která je dána licenční politikou a násobně zvyšuje náklady

s počtem přibývajících let. Vhledem k tomu, že společnost LMC je na search enginu přímo závislá, **není pro ni z dlouhodobého hlediska varianta (V4) realizovaná formou outsourcingu společností Jyxo vhodná**.

- Varianta (V3) nabízená společností SearchBlox je nejvýhodnější z pohledu její ceny, ale nesplňuje kritéria společnosti LMC ohledně požadovaného rozsahu funkčnosti, a to především v podpoře českého jazyka a možnosti využít vlastní algoritmy pro strukturovaná kritéria. **Variantu (V3) od společnosti SearchBlox tedy nelze doporučit**.
- Varianta (V1) nabízená společností Jyxo, je prakticky shodná s variantu outsourcingu, ale search engine je provozován na vlastním HW vybavení. Slabší stránkou tohoto řešení je pouze rozsah požadované funkčnosti a nutnost jejího doimplementování. Náklady na tuto variantu jsou již srovnatelné s ostatními variantami, ale zůstávají zde neustálé měsíční poplatky za provoz a závislost na poskytovateli řešení při dalším vývoji. **Proto není varianta (V1) považována za zcela vhodné řešení**.
- Poslední variantou (V2), kdy lze zakoupit již kompletní řešení, je aplikace xMP od společnosti Actonomy. Aplikace xMP poskytuje značný rozsah požadované funkčnosti včetně nástroje pro generovaná vlastních kategorií a možnosti využití různých algoritmů pro strukturovaná kritéria, ale zatím plně nepodporuje český jazyk. Proto i přes příznivý poměr ceny a výkonu **není varianta (V2) realizovaná společností Actonomy zcela vhodná**.
- Závěrečné hodnocení je věnováno variantě (V5), která nabízí implementaci technologie Apache Lucene na HW vybavení společnosti LMC. Tato společnost má s realizací obdobných projektů již několikaleté zkušenosti a je ochotná poskytnout své dosavadní know-how i rozsáhlé slovníky generované tiskovými médii nejen v ČR.

  Podle výsledků metrik nabízí tato varianta nejvíce z rozsahu požadované funkčnosti, která se také projevuje v počáteční náročnosti implementace a vyšších nákladech, které se však v průběhu provozu usměrní pouze na náklady spojené se správou, které jsou mnohonásobně nižší než u zakoupených řešení. Další výhodou tohoto řešení je nesporně znalost search enginu s možností dalšího vývoje, včetně jeho využití v rozdílných aplikacích, a to přímo týmem programátorů v LMC při

značně nižších interních nákladech.

Pokud se podíváme na výkonnost tohoto řešení uvedenou v kapitole *Výkonnost jednotlivých variant*, tak splňuje požadavky společnosti LMC na průměrnou dobu odezvy stanovenou na < 1 vteřinu, a to i v případě, kdy je v dotazech více jak 25 kritérií a představují zátěž 100 paralelních uživatelů. Z těchto důvodů byla **varianta (V5) využívající open source technologii Apache Lucene implementována s pomocí společnosti Conlegere, doporučena jako optimální řešení pro společnost LMC.**

9 Závěr

Cílem této práce bylo doporučit optimální řešení vyhledávácího nástroje (search enginu) pro společnost LMC, která je provozovatelem pracovních serverů.

Dílčí cíle:

- Popsat současné využití search enginu ve společnosti LMC.
- Představit principy současných web search enginů.
- Provést analýzu současných řešení na trhu a vybrat taková, která jsou vhodná pro vyhledávání na pracovních serverech.
- Zhodnotit jednotlivé varianty search enginů na základě předem stanovených metrik a stanovit optimální řešení pro společnost LMC.

První kapitola nazvaná „Úvod“ byla zaměřena na obecný úvod do problematiky vyhledávání na Internetu. Dále byly vysvětleny pojmy e-recruitment, elektronický trh práce a pracovní servery. Závěr této kapitoly byl věnován příkladům služeb, které dnes nabízejí provozovatelé pracovních serverů a mezi které patří i různé typy search enginů.

Ve druhé kapitole s názvem „Popis současného řešení “ byla představena společnost LMC a různé typy vyhledávání, které ve svých službách nabízí. V další části byla popsána architektura a principy současně provozovaných aplikací „Search“ a „Agenti“. V poslední části jsou zmíněny nedostatky současného řešení.

Ve třetí kapitole nazvané „Požadavky společnosti LMC na nový search engine“ byly specifikovány požadavky, které jsou následně zohledněny při výběru nového search enginu. Mezi tyto požadavky patří podporované technologie, využitelnost v rámci různých aplikací, výkonnost, management, rozsah podporované funkčnosti a náklady na implementaci řešení.

Čtvrtá kapitola nazvaná „Koncepce search enginů“ byla věnována architektuře současných web search enginů a následnému popisu procesů probíhajících při indexování a vyhledávání dokumentů.

Obsahem páté kapitoly nazvané „Možnosti řešení search enginů“ bylo provést analýzu současně nabízených search enginů na trhu a vybrat z nich ty, které jsou vhodné pro vyhledávání na pracovních serverech. Vybrané varianty řešení reprezentované jejich

dodavateli, byly následně rozděleny do třech kategorií: *Zakoupení již hotového řešení*, *Forma pronájmu*, *Využití open source technologií.*

V šesté kapitole nazvané „Stanovení metrik“ byly představeny základní charakteristiky metrik a následně byly definovány metriky, které byly využity při výběru vhodného řešení v kapitole *Porovnání a návrh vhodného řešení.*
Zvolené metriky:

- metrika náročnosti implementace a správy search enginu,
- metrika rozsahu požadované funkčnosti,
- metrika výkonnosti search enginu,
- metrika celkových nákladů.

V sedmé kapitole nazvané „Výkonnost search enginů“ byly stanoveny požadavky a nastíněn postup testování výkonnosti různých variant search enginů. Samotný test nebyl z důvodu náročnosti na vstupní požadavky proveden, a proto jsou v druhé části této kapitoly uvedeny alespoň některé výsledky testů získané od poskytovatelů řešení.

V poslední kapitole nazvané „Porovnání a návrh vhodného řešení“ byla nejprve provedena aplikace stanovených metrik na jednotlivé varianty řešení. Výstupy z těchto metrik byly následně spolu s požadavky specifikovanými v kapitole *Požadavky společnosti LMC na nový search engine* zohledněny v závěrečné části, kde byly porovnány výhody a nevýhody jednotlivých variant a navrženo optimální řešení. Pro společnost LMC bylo navrženo optimální řešení s využitím open source technologie Apache Lucene implementované společností Conlegere.

10 Literatura

[Ucen_01] Učeň, Pavel : **Metriky v informatice.** Jak objektivně zjistit přínos informačního systému. Grada Praha, 2001.

[Hlavenka_04] Hlavenka, J. : **Mistrovství ve vyhledávání na Internetu.** Computer Press Brno, 2004.

[Zhou_06 www] Zhou, D. P. : **Beef up Web search applications with Lucene**.
http://www.ibm.com/developerworks/web/library/wa-lucene2

[Brin Page www] Brin, Sergey; Page, Lawrence : **The Anatomy of a Large-Scale Hypertextual Web Search Engine.**
http://infolab.stanford.edu/~backrub/google.html

[RFC1945_96 www] Fielding, R; Frystyk, H. : **RFC 1945 - Hypertext Transfer Protocol.**
http://www.faqs.org/rfcs/rfc1945.html

[RFC1945_96 www] Fielding, R; Frystyk, H. : **RFC 1945 - Hypertext Transfer Protocol.**
http://www.faqs.org/rfcs/rfc1945.html

[Biroscak_06] Biroščák, R.: **Kontextové prehľadávanie textu**. Bakalářská práce, ČVUT Praha, 2006.
http://dce.felk.cvut.cz/dolezilkova/diplomky/2006/bp_2006_biroscak_rastislav/bp_2006_Biroscak_Rastislav.pdf

[Porter_80 www] Porter, M. F. : **An algorithm for suffix stripping**.
http://tartarus.org/martin/PorterStemmer/def.txt

[Lucene www] Lucene.apache.org **: Stránky v angličtině věnované open source projektu Apache Lucene**.
http://lucene.apache.org

[Lacvik_07 www] Lacvik, M.: **Vyhl'adávanie informácií**. Slovenská technická univerzita v Bratislavě, 2007
http://www.laclavik.net/publications/vi_laclavik.pdf

[Jobs.cz www] Jobs.cz : **Odborně zaměřený pracovní server**.
http://www.jobs.cz

11 Použité termíny

Agent – robot pro automatizované vyhledávání PD

API (application programming interface) – představuje určité třídy, rozhraní a metody, pomocí nichž lze s aplikaci, která API nabízí, pracovat.

Brand – představuje jeden vybraný pracovní server z VPV

CV- Curriculum Vitae představuje životopis uživatele

DB - databáze

E-recruitment - získávání pracovníků pomocí elektronických sítí

Framework - je určitý rámec předdefinovaných tříd, rozhraní, metod a postupů, které slouží jako podpora při programování nových projektů.

HTTP – (Hypertext Transfer Protocol) je jednoduchý, bezestavový protokol postavený na principu požadavek/odpověď sloužící pro přenos informací mezi webovým serverem a aplikací.

Index – představuje úložiště dat, které již jsou převedeny do podoby, nad kterou lze provádět rychlé vyhledávání.

JD – Job description představuje charakteristiku volné pozice.

Metrika (metrika IS/ICT) – měřitelná veličina. Metrika IS/ICT je přesně vymezený ukazatel nebo hodnotící kriterium, který je používán k hodnocení úrovně efektivnosti či

jakosti konkrétní oblasti IS/ICT k hodnocení podnikového výkonu nebo k hodnocení

úrovně podpory podnikového procesu prostředky IS/ICT.

ONREA (online recruitment europe applications) – společenství pracovních serverů v Evropě.

Outsourcing (Outside Resource Using) – podstatou outsourcingu je vytěsňování či vyčleňování určitých podnikových činností z podniku a jejich zabezpečení u externího dodavatele. Outsourcing tedy představuje využití vnějších zdrojů.

PD – prezentační dokument, který může být typu CV nebo JD.

SAS – (Serial Attached SCSI) je inovovaná generace rozhraní SCSI zvyšující výkonnost ukládacích systémů a zlepšující jejich dostupnost.

Search engine – aplikace, která zajišťuje zpracování vstupního dotazu a nalezení odpovídajících dokumentů.

VPV – veřejná prezentační vrstva se skládá z veřejně dostupných pracovních serverů bez nutnosti autentizace. Např. servery jobs.cz, prace.cz, hotjobs.cz, topjobs.sk

Webové služby – představují komunikační rozhraní mezi aplikacemi.

Wokflow – představuje specifický proces, který je rozdělen na jednotlivé činnosti a jejich vazby. Příkladem může být proces nalezení nových uchazečů o zaměstnání, a to včetně jejich vyhledání, pozvání na pohovor, schválení, přijmutí a dalšího rozvoje.

12 Přílohy

12.1 Různé implementace a rozšíření pro Apache Lucene

Následující projekty, které implementují a rozšiřují funkčnost Apache Lucene knihovny, jsou zde uvedeny pouze jako inspirativní s tím, že některé části mohou být využity při případné implementaci.

- Solr - http://lucene.apache.org/solr je open source enterprise search server založený na Lucene Java search knihovně, s XML/HTTP a JSON API, zvýrazněním nalezených klíčových slov (highlightingem) a web administračním rozhraním.
- Nutch - http://lucene.apache.org/nutch je open source web search založený na Lucene. Má vlastní crawler a analyzátory pro indexování různých typů dokumentů

- Kneobase - http://www.kneobase.com je enterprise search engine postavený na Lucene a Spring frameworku.

12.2 Další možná řešení search enginu

- Zde uvádím přehled search enginů, které nebyly z důvodu nedostatku informací zařazeny do hodnocení.
- Morfeo – (www.morfeo.cz) – je komerční search engine provozovaný společností NetCentrum s.r.o. který využívá technologii Sherlock.
- Sherlock search engine – (www.ucw.cz/holmes) je open source projekt realizující textový search engine napsaný v jazyce C.
- Egother – (www.egothor.org) - je open source projekt ve formě knihovny napsaný v jazyce Java založený především pro akademické účely.
- FAST – (www.fastsearch.com) je komerční, komplexní řešení search enginu, které získalo již několik prestižních ocenění např. Best Enterprise Search Engine 2007