Posudek vedoucího diplomové práce

Martin Košalko: Alternativní vyhledávač systému EGOTHOR

Cílem práce bylo navrhnout a implementovat alternativní vyhledávací modul pro vyhledávač Egothor. Navržené řešení implementuje vektorový vyhledávač s tím, že je možné vyhodnocovat dotazy pomocí obou způsobů zároveň a na základě modulu pro optimální vyhledávání kombinovat výsledky dohromady. Hlavními částmi práce jsou ty, jež popisují zvolená řešení vyhledávače FRC, tedy kapitoly 4-8. Kapitola 4 je věnována jednak obecnému popisu vektorového modelu, ale především odvození návrhu indexových struktur, potřebných pro činnost výsledného DIS, spolupracujícího s Egothorem. Další kapitoly popisují modulární strukturu řešení, zvolené algoritmy pro konverze dotazů mezi jednotlivými paradigmaty a spojování odpovědí z vyhledávačů.

Práce implementuje čtyři rozdílné způsoby výpočtu míry podobnosti mezi dotazem a dokumenty. V testovací aplikaci je možné použít paralelně všechny čtyři modely výpočtu podobnosti, ale při přepočítávání koeficientů pro použité vektorové vyhledávače na základě zpětné vazby mi vycházely všechny koeficienty vždy stejně. To by odpovídalo tomu, že všechny čtyři našly vždy stejné dokumenty a proto je použitý a v práci popsaný algoritmus mezi nimi nedokázal kvalitativně rozlišit. Zajímalo by mne, zda existuje nějaký dotaz a jeho ohodnocení, které by vedlo k rozdílným oceněním použitých DIS, nebo zda se jedná o nějakou chybu v konfiguraci.

Vektorové dotazy se nad indexem vyhodnocují v několika krocích, což je dáno využitím původního indexu a algoritmů obsažených v Egothoru pro předvýběr dokumentů, pro které se podobnost počítá. Práce uvádí čas potřebný k vyhodnocení vektorového dotazu o zhruba 35% vyšší, než je čas potřebný k vyhodnocení dotazu boolského. Nikde není uvedeno, kolik procent navíc vyžadují doplňkové indexy. Demonstrační data na CD napovídají, že se jedná o zhruba 50% objemu původního indexu. Nebylo možné se nárůstu vyhnout, případně jej minimalizovat?

Kapitole 8, popisující rozhraní modulů mi nebylo příliš jasné, jak je to s rozhraním I_4. V posledním odstavci na straně 53 je uvedeno, že je nejširší a zároveň je uvedeno, že rozhraní I_5 je jeho nadmnožinou. Samotné rozhraní I_5 pro obsluhu distribučního modulu je podle kapitoly 8.1 pouze grafické. Je tomu skutečně tak, nebo je procedurální a aplikace je pouze jeho nadstavbou?

Uživatelská příručka v příloze A by mohla být podrobnější. Například jména tříd, implementujících konvertory dotazů, je možné najít jen v kapitole 7.3. U popisu konfigurace vyhledávače o nich není žádná zmínka. Samotná aplikace v Javě je ve stádiu prototypu a její uživatelská přívětivost není nejlepší. O to více bych uvítal podrobný popis všech jejích funkcí. Pokud se například některý z vyhledávačů deaktivuje stiskem tlačítka [*Deactivate*], systém zapomene cestu ke třídě, a jeho opětovná aktivace se mi již nepodařila. Konfiguraci použitých vyhledávačů je možné uložit a později načíst, ale po každém načtení je nutné vyhledávače aktivovat, což obnáší opětovné určení adresáře s indexem. To mi přijde zbytečné, stejně jako nutnost vždy znovu při aktivaci generovat doplňková data pro FRC z indexu Egothoru.

Celkově se domnívám, že práce splnila požadavky kladené na práci diplomovou a přes své výše uvedené výhrady ji doporučuji k obhajobě.

V Praze dne 16. 5. 2006

RNDr. Michal Kopecký, Ph.D.
KSI MFF UK

Dotazy (na druhé straně)