Oponentský posudek diplomové práce

Název DP: **Mathematical search engine**
Diplomant: **Jozef Mišutka**

Obsah práce:

Předmětem diplomové práce je návrh architektury a implementace vyhledávače matematických dokumentů podle matematických vzorců-formulí v nich obsažených. Matematické vyhledávání je implementováno jako rozšíření stávajícího fulltextového vyhledávače Egothor (vyvinutého dr. Galambošem na MFF). Autor v úvodní kapitole popisuje cíle práce, stručně cituje existující práci v oblasti vyhledávání (podle) matematických výrazů. Ve druhé kapitole je obsáhle diskutována celá problematika a dílčí přístupy k vyhledávání - zde je také navrženo konkrétní ucelené řešení zamýšleného vyhledávacího stroje, včetně indexování. Třetí kapitola se věnuje indexování matematických formulí a čtvrtá jejich vyhledávání. V páté kapitole autor popisuje implementaci prototypu. Šestá kapitola obsahuje experimentální vyhodnocení vyhledávacího stroje na reálných kolekcích dat. Poslední kapitola shrnuje práci a diskutuje výsledky.

Hodnocení:

Diplomová práce představuje ucelené softwarové dílo a jeho pečlivou analýzu. Je třeba ocenit povahu díla navazujícího na předchozí práci, realizovanou ve vyhledávači Egothor. Diplomových prací této povahy je bohužel poměrně málo. V práci není jasně vyznačeno, zda obsahuje původní dílčí výsledky i v teoretické části, za původní přínos tedy považuji návrh architektury a implementaci. Kromě pečlivé analýzy jednotlivých kroků oceňuji i experimentální část, kde autor musel vymyslet metodiku vyhodnocení kvality vyhledávání podle matematických výrazů, která v současné době neexistuje v obecně přijímané podobě. Na druhou stranu, v experimentech by čtenáři v orientaci pomohlo srovnání s jinými vyhledávacími stroji, ať je to zmíněný MathWebSearch, nebo prostý fulltextový vyhledávač. Zajímavé by byly rovněž časové výsledky, ať již rychlost indexování, tak rychlost vyhledávání. Celkově je vidět, že autor odvedl velký kus práce.

Text práce je psán v angličtině, která je na velmi dobré úrovni. Formálně práce splňuje všechny požadavky.

Podrobnější připomínky, poznámky:

1) Ačkoliv je angličtina dobrá, někde se vyskytuje český-slovenský slovosled a některá špatně přeložena slova (např. analyse místo analysis, atd).

2) Celá myšlenka indexování je založena na fulltextovém přístupu, potažmo indexování termů (zde tokenů) invertovanými seznamy. Zajímalo by mě, zda autor uvažoval o

zcela jiném přístupu, který podporuje samotný model generalizačních pravidel. Jednalo by se o (obecně nevyváženou) stromovou strukturu podobnou např. R-stromu, kde místo minimálních obdélníků ořezávajících vyhledávací prostor by sloužilo zobecněné pravidlo ořezávající „prostor výrazů". Podstrom uzlu by rekurzivně obsahoval spezializované výrazy spadající do šablony nadvýrazu. Vedlejším produktem tohoto přístupu by bylo vytvoření shluků podobných výrazů (a potažmo dokumentů výrazy obsahující) v listech stromu, což by se dalo využít např. pro kategorizaci dokumentů nebo pro jiné „datamining" účely.

Závěr:

Práce splnila zadání, autor v ní tvůrčím způsobem navázal na předchozí práce na MFF. Práci doporučuji k obhajobě.

V Praze dne 1. září 2007

Doc. RNDr. Tomáš Skopal, Ph.D.
oponent