Posudek diplomové práce

Jméno: Michal Král
"Dotazování databází a webu"

Cílem práce bylo navrhnout prototyp, který umožní rozšířit dotazování nad tradičními databázemi (RDBMS) o dotazování vyhledávacích strojů na webu (VSW). Dalším úkolem bylo provést experimenty na netriviálních datech.

Práce je členěna do tří částí. První popisuje stávající techniky rozšiřování RDBMS o IR systémy, druhá technické aspekty takového rozšíření, a následně je popsána architektura navrhovaného prototypu spolu s požadovanými experimenty.

Jádrem prototypu je konstrukce tří funkcí www_rank, www_near a www_best_address. Ty jsou vypočítávány na základě údaju z VSW. Do RDBMS jsou importovány přes rozhraní nad fiktivními tabulkami, což umožňuje jejich bezproblémové napojení do relační technologie.

Podle mého názoru se autor zaměřil pouze na výchozí reference ze zadaní, ale již méně hledal souvislosti s okolními technologiemi respektive s oborem vyhledávacích strojů. Nejmarkantněji je to patrné při realizaci funkce www_best_address, kterou bylo možné konstruovat širokou škálou základních algoritmů, například HITS, Clever, RankFile ad. Místo toho byl zvolen vlastní postup dedukce nejlepší webové adresy pro daný řetězec. Zvolený postup je jistě inovativní a rychle vypočitatelný, ale dává výsledek, který ani nemusí být skutečnou webovou adresou. Navíc je funkční pouze v situaci kdy přijmeme premisu, že zanořenější URI obsahuje zpřesnění nějakého tématu nadřazeného URI. To je postup, který muže dobře fungovat pouze pro malé weby, nikoliv obecně, kdy je toto zpřesňování často vedeno horizontálně přes několik domén v souladu se strukturou odkazu. V této části práce mi tedy chybí zduvodnění zvoleného postupu a jeho diskuze v rámci známých postupů.

Další vážnější chybou je neujasněnost terminologie. Na straně 5 hovoří o "používanosti", zatímco na straně 36 přechází na termín "důležitost". Následně vychází z předpokladu, že důležitost odpovídá četnosti. To bych pokládal bez dalšího objasnění jako chybné. Alespoň s ohledem na známé proměnné v oblasti IR: tf a idf.

V neposlední řadě autor uvádí nepravdivé informace. Například na straně 40 je uvedeno, že znak plus v URL znamená, že výskyt slova je na stránce povinný. To je nepravdivé tvrzení, protože uvedené plus znamená pouze znak mezery. Znak plus by musel být kódován jako %2B. Je pak otázkou, zda potom práce pracuje s hodnotami, které skutečně popisuje a předpokládá (do URL kóduje dotazy do VSW).

V závěru je uvedeno, že funkce www_best_address vrací relevantní hodnoty. Osobně se domnívám, že termín "relevantní" chápu diametrálně odlišně než autor.

Rozsáhlejší zdůvodnění by si zasloužily i některé konkrétní výroky. Například tvrzení, že "prohledávat internet úplně bez využití DB vlastně nelze a vždy se tak bude jednat o DSQ" (str. 10), bych s ohledem na okolní kontext spíše chápal jako "prohledávat internet úplně bez využití datových struktur vlastně nelze", což je bezpochyby pravda. Není mi pak jasné jaký rozdíl klade autor mezi DB a datovou strukturu.

Na straně 13 je uvedeno, že "v případě simulace pomocí WebPage bude pravděpodobně každý odkaz vrácen právě jednou". Je otázkou za jakých podmínek tento výrok platí a jak je kvantifikován termín "pravděpodobně". Pokud bychom veškerý web napsali na jednu stránku, tak bude zřejmě vrácena mnohokrát.

Přesnější vyjadřování by si zasloužilo i referencování tří konstruovaných funkcí. V práci nebyly očíslovány a přesto je autor referencuje výrazy "první", "druhá" a "třetí" (str. 10-13). Toto označení

se pak prolíná se slovním referencováním ukázkových dotazů, což na stránce 13 vede ke třem zcela odlišným interpretacím výrazu "druhý dotaz":

1. ten bez WebPage
2. ten z bloku o druhé funkci
3. ten druhý v tomto bloku s WebPage

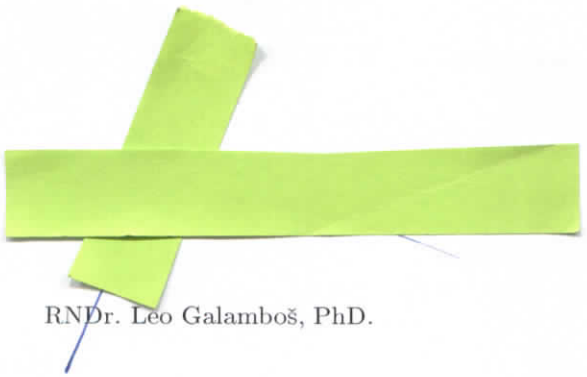
Na straně 14 autor uvádí, že "do výbavy vyhledávačů patří pouze možnost omezit stránky odkazující na určitou adresu". V obecném znění (bez vymezení množiny vyhledávačů) je tato věta nepravdivá. Podobně nejasně působí i věta o tom, že VSW "neumožňují využít údajů uložených přímo v DB". Domnívám se, že to právě činí a nabízí světu přes rozličná API.

Práci rovněž doprovází tajemné informace, například "podporovaná je verze 1.4, s verzí 1.5 bývají problémy a pokud to jde, je lépe se jim vyhnout". Je tedy podporována jen verze 1.4 anebo jsou nějaké problémy s něčím co má fungovat? S čím konkrétně? Na straně 32 je zase uvedeno, že "roli sehrála schopnost nastavení výsledků vyhledávání pomocí *změny adresy stránky*", což lze jen obtížně interpretovat. Na straně 43 by lépe místo sdělení "testy poběží na pevném připojení" posloužil konkrétní *traceroute*.

Z dalších nepřesností: java.net.URL není knihovna (str. 31), proč musíme pracovat na intervalu 0-100000 (str. 42).

Z typografického hlediska pozitivně vítám, že je práce vysázena v TEXu. Na druhou stranu trpí mnohými typografickými chybami (oboustranná sazba v jednostránkovém tisku atp.). Je rovněž škoda, že práce obsahuje hrubé gramatické chyby (shoda podmětu s přísudkem, str. 39: "údaje byli dobrým kandidátem").

Celkově práce splnila zamýšlené cíle a splňuje nároky kladené na diplomové práce. Proto ji doporučuji k obhajobě.



RNDr. Leo Galamboš, PhD.

V Praze, dne 29.4.2007