Oponentský posudek k diplomové práci

V. Kulvait: **Crawlování na Webu**

Práce p. Kulvaita se věnuje srovnávání různých metod crawlování. Výsledky, plynoucí z jednotlivých měření ukazují zajímavé a někdy až nečekané souvislosti.

Autor v úvodu vymezil téma, kterému se bude věnovat. Následuje teoretický úvod do této tematiky, obohacen o fakta a informace týkající se největších vyhledávačů (např. počet stránek, se kterými pracují). Po kapitole věnované struktuře webu a přístupům vyhledávání následuje část, zabývající se samotnými crawlery a strategiemi crawlování. Další kapitola popisuje dvě míry důležitosti – PageRank a K-Rank, které bude autor detailněji zkoumat a porovnávat. Po poslední teoretické kapitole, vysvětlující Kendallovo τ, následuje popis použitých datových struktur a vysvětlení implementovaných algoritmů. Následuje popis experimentů a vyhodnocení výsledků. V posledních kapitolách je celkový přehled s popisem vlastní autorovy práce, závěr s diskuzí a zamyšlení nad budoucím možným pokračováním.

Vstupní data byla rozdělena do tří souborů, reprezentujících kolekci stránek domény .cz, .uk a .edu, stažených z Webu, takže se pracovalo se skutečnými daty. Autor počítal hodnocení K-Rankem na matici Webu, která byla až o třetinu menší než matice, kde se počítal PageRank, proto výsledky nelze úplně srovnávat. Nicméně i z těchto výstupu vyplývá, že K-Rank rozkládá důležitost rovnoměrněji než PageRank. Jeho uspořádání vzhledem k celkové agregované důležitosti je stabilnější než u PageRanku.

Mé výhrady směřují zejména k formální stránce práce:

- Grafická úprava. Předložky v textu nejsou vázány na následující slovo, proto se často stává, že je třeba dvojice slov „V nové“ rozdělena ta, že je „V“ na konci řádku.
- Na str. 5 je uvedeno, kolik odkazů najdou největší vyhledávače na dotaz „and“ a pak následuje věta: „Největší vyhledávače tak pravděpodobně spravují kolekce, které čítají přes 20 000 000 000 stránek.“ Odkud je tento závěr?
- Některé termíny (např. PageRank) jsou používány v textu před jejích definicí
- Občas je vyjadřování nepřesné – kap. 4.7. strategie si vede dobře – jde o čas, pořadí nacrawlovaných stránek nebo výslednou množinu stránek?

Celkově považuji práci p. Kulvaita za kvalitní, autor navrhl a implementoval algoritmy pro crawlování, čímž zcela splnil zadání. Proto s ohledem na výše uvedená fakta doporučuji tuto práci k obhájení.