

## Posudek diplomové práce

Peter Irikovský:

“Graph algorithms in text retrieval”

Cílem práce bylo prozkoumání možností využití grafových algoritmů v oblasti vyhledávání informací (IR).

V první části, v kapitolách 1-4, je podán obsáhlý přehled všech základních grafových algoritmů a to včetně nástinu jejich využití v rámci IR. Následně se autor zabývá existujícími algoritmy přímo z prostředí Webu a to zejména algoritmy PageRank, HITS a také algoritmy pro odhalování komunit.

Kromě rešeršní části autor implementoval i algoritmus pro výpočet PageRank-u a demonstroval pak jeho použití na reálných datech. Při provozování na běžném PC je tento program primárně nasaditelný na kolekce v řádu desítek miliónu dokumentů. Na druhou stranu, při doimplementování potřebných technik z rešeršní části práce je možné daný rozsah zvětšit.

Z teoretické části je patrné, že má autor dostatečný přehled v popisovaných technologiích. Na druhou stranu bych doporučoval určitá upřesnění. Například:

- kapitola 5.1.2: vzorec pro výpočet PageRank-u ve formě  $P = H \cdot P$  by s ohledem na předchozí text měl spíše znít  $P = D \cdot H \cdot P + \frac{1-D}{n} \cdot (1, 1, \dots, 1)^T$
- kapitola 5.1.4.6: z textu mi není zcela jasné jakým způsobem efektivně měnit velikost buněk pole po vektor PageRank-u v případě využití Pack7 formátu
- kapitola 5.2.7: neprůkaznou tabulku by bylo vhodnější nahradit log · log grafem

Jazykové zpracování v anglickém jazyce hodnotím jako nadstandardní a je mu jen s obtížemi co vytknout. Chyby jsou okrajového charakteru, například:

- kapitola 3.1: “A graph is *a* called a tree...”
- kapitola 5.1.4.5: citace literatury “f30” a “f31”
- kapitola 5.2.1: “However, *is* the number is stored... we would still need...”

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