Posudek oponenta diplomové práce

Vladimír SADLOŇ

Detekce duplicit v rozsáhlých webových bázích dat

Cílem práce bylo prostudovat metody detekce plagiátů textových dokumentů a implementovat vhodné rozšíření do vyhledávacího stroje *Egothor 2*. Práce nejprve definuje problém plagiátorství, následně shrnuje nastudované používané metody, popisuje prostředí vyhledávacího stroje a implementaci vybrané metody.

Text je poměrně přehledný a dobře čitelný. Nicméně občas je možné najít překlepy a drobné chyby ve vzorcích a značení. Například:

- V sekci 2.2 na straně 4 má být dokument Q členěn na části q_j, nikoli q_i. Horní hranice indexů i a j jsou zavedeny jako velká písmena N a M, ale ve vzorci 2.1 je jako horní hranice použito malé písmeno n.
- Vzorec 3.2 na straně 14 pro sublineární škálování zřejmě předpokládá, že frekvence termů $tf_{i,j}$ jsou celá čísla, tedy počty výskytů. Bylo by lepší význam pojmu v rámci práce explicitně uvést, protože – pokud by byla frekvence chápána jako podíl počtu výskytů a velikosti dokumentu – vycházely by logaritmy vesměs záporné. Rovněž není z textu jasné, co za hodnotu je x ve vzorci pro případ, kdy je $tf_{i,j}$ nulová (předpokládám, že záporná být nemůže, i když definice počítá i s takovou možností). Předpokládám, že $0 \leq x \leq 1$, nejspíše $x = 0$.

V přehledové části je uveden reprezentativní výčet nastudovaných používaných metod. Autor hodnotí vhodnost jejich použití v zamýšleném kontextu, a z možných variant nakonec – s ohledem na požadavky – vybrána metoda tzv. šindelování, která (teoreticky) slibuje čas zpracování jednoho dokumentu, který je nezávislý na velikosti referenční kolekce dokumentů.

Experimentálnímu zhodnocení práce autor věnuje více než třetinu práce, což značně převyšuje obvyklé úsilí u prací konkurenčních. V rámci ní jsou testovány různé parametry porovnávání jako jsou délky úseků, které šindele tvoří a počty permutací, které se testují. Reálně dosažené výsledky ukazují, že výkon aplikace výrazně převyšuje konkurenční nástroje, alespoň ty, které bylo možné s implementovanou verzí porovnat.

Detekce plagiátů je subjektivně měřeno rychlá – odezva je v rámci sekund – a rozdíly ve výsledcích jsou pro originální práci a plagiát velmi výrazné, takže je ve výsledkové listině plagiát poměrně snadné identifikovat. Domnívám se, že by bylo vhodné uvažovat o zprovoznění detektoru plagiátů přímo v rámci fakulty nebo katedry, aby mohli vyučující texty prací kontrolovat již před jejich odevzdáním.

Je otázka, nakolik je odhad doby vykonávání dotazu, odvozovaný v sekci 5.4 na straně 35-36, a založený na tom, že šindele se v kolekci dokumentů objevují jen jednou (respektive max. konstantně krát nezávisle na množství dokumentů) skutečně oprávněný na opravdu velkých kolekcích. Řekl bych, že počet výskytů konkrétního šindele bude na počtu dokumentů nějak lineárně záviset, byť s malou multiplikativní konstantou, takže pro kolekce obsahující tisíce dokumentů, jako je tomu v případě stávajících testů, je to zanedbatelné, ale pro kolekce, sesbírané z celého Internetu a obsahující miliardy dokumentů tomu může být jinak.

Přes výše uvedené drobné připomínky práce splnila jak zadání, tak požadavky, kladené na práci diplomovou. Doporučuji ji proto k obhájení.

V Praze dne 18. 1. 2011

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