POSUDEK DIPLOMOVÉ PRÁCE

VLADIMÍR SADLOŇ: DETEKCE DUPLICIT V ROZSÁHLÝCH WEBOVÝCH BÁZÍCH DAT

Cílem práce bylo vytvořit prakticky použitelný systém pro detekci duplicit, který využije platformy existujícího webového vyhledávacího stroje (WVS). Dále mělo dojít k prostudování existujících postupů pro vyhledávání plagiátů, popsání datových toků v rámci celého procesu, vyhodnocení I/O nároků, a rovněž mělo dojít ke srovnání s jinými strategiemi.

Vyhledávání plagiátů nelze zjednodušeně zaměnit za vyhledávání „podobných dokumentů“, neboť takový proces typicky nezohledňuje strukturu dokumentu a zaměřuje se toliko na jeho reprezentaci (jako celku) v rámci vyhledávacího stroje. V tomto případě by se jednalo o vektorovou reprezentaci s ohledem na volbu WVS Egothor2. Proto bylo stěžejním úkolem navržení takového postupu, který dovoluje efektivně vyhledávat podobné/shodné úseky dokumentů s využitím již existujících procesů a datových struktur reálného WVS.

Práce staví na dosažených výsledcích diplomové práce Mgr. Kateřiny Dufkové (MFF UK 2008). Přidává však podrobnější náhled na produkčně použitelnou implementaci. Výsledné softwarové dílo nabízí přibližně 5x rychlejší zpracování dotazu, včetně možnosti zpracování dokumentových bází většího rozsahu. Původní práce Mgr. Kateřiny Dufkové obsahovala prototypové řešení, které bylo navíc omezeno velikostí vnitřní paměti.

Kromě toho nedochází k narušení interních procesů a datových struktur WVS. Je tak možné provádět běžné operace s indexem, které reflektují změnu indexované báze dokumentů: přidávání nových dokumentů, změny v dokumentech a jejich odstraňování. Jinými slovy, běžný provoz WVS není nikterak narušen a přitom je možné efektivně vyhledávat shodné části vstupního dokumentu vůči zaindexované množině dokumentů. Při využití webové aplikace Bc. Miroslava Tamáše (Bakalářská práce, MFF UK 2009) je proto možné provádět vyhledávání plagiátů s využitím libovolné podmnožiny dostupných indexů, například Wikipedia, stránky v CZ TLD, báze jednotlivých fakult a univerzit, apod.

V samotném textu práce se autor nejdříve zabýval definicí pojmu „plagiát“ a stanovením cílů výsledné implementace (str. 1-9). Poté popsal v současnosti používané postupy pro vyhledávání plagiátů (str. 10-19). Dále se zabýval volbou vhodného algoritmického řešení (str. 25-39) a popisem interních procesů WVS Egothor2 (str. 20-24) tak, aby mohl navrhnout vhodné invazivní řešení pro daný WVS (str. 40-49). Realizované řešení bylo podrobeno testování a ověření vlastností na reálných datech (str. 50-63). V neposlední řadě došlo i k srovnání vůči existujícím systémům pro odhalování plagiátů (str. 64-68).

S ohledem na splnění všech cílů práce, včetně vytvoření produkčně použitelného softwarového díla, ji doporučuji k obhajobě a jako takovou ji uznat i jako práci diplomovou.

V Praze, dne 18.12.2011 RNDr. Leo Galamboš, Ph.D.