Posudek na diplomovou práci

Jan Sochna

Systém pro sběr XML dat a metadat z internetu

Cílem práce je implementace systému, který by umožnil systematické pořizování vzorků dat z rodiny XML jejich stahováním z Internetu, a to především pro účely statistické analýzy těchto dat. Práce navazuje na studentský projekt Analyzer, jehož spoluřešitelem byl autor práce.

Text práce nejprve popisuje základní problémy, s nimiž se sběr dat z Internetu potýká. Úroveň tohoto popisu je přiměřená jeho účelu, chybí ovšem zřetelnější oddělení problémů specifických pro XML, jejichž řešení je předmětem této práce, od problémů všeobecných, u nichž je řešení pouze přejato.

Další kapitola popisuje existující systémy pro sběr dat z Internetu, a to jak systémy přímo zaměřené na XML, tak systémy obecné, které by mohly být pro XML přizpůsobeny. Zatímco systémů orientovaných na XML existuje velmi málo, crawlerů pro HTML existuje obrovské množství a je otázka, zda je autorův výběr dostatečně reprezentativní, protože v textu není uvedeno, která kriteria pro výběr autor zvolil. V rámci daného výběru pak autor přichází k víceméně jednoznačnému závěru, že nejvhodnějším řešením je úprava existujícího crawleru Apache Nutch.

Ve čtvrté kapitole je popsána samotná implementace systému, přesněji řečeno, seznam úprav, které autor provedl jak v Apache Nutch, tak v projektu Analyzer. Nutnost těchto úprav byla ve většině případů zdůvodněna již v druhé kapitole, čtenář neznalý detailů systémů Nutch a Analyzer ovšem poněkud ztrácí celkový přehled a nemá jistotu, že zvolený způsob úprav je optimální.

Závěrečné kapitoly pak především popisují výsledky ve formě různých statistických údajů, mimo jiné srovnávajících chování upraveného systému s původním systémem Nutch. Ze striktně statistického pohledu tyto údaje nejsou příliš spolehlivé, neboť jde o výsledky testů prováděných v průběhu několika měsíců a vzhledem k proměnlivosti Internetu nejsou opakovatelné. V rámci možností však tyto výsledky ukazují, že autorem provedené úpravy výrazně zlepšují použitelnost systému při získávání dat z rodiny XML. Je ovšem škoda, že se autor nepokusil změřit přínos jednotlivých provedených úprav individuálně.

Až na výše uvedené nedostatky je text práce dobře členěn a zpracován s požadovanou úrovní přesnosti, takže z tohoto hlediska vyhovuje nárokům na diplomovou práci.

Přiložený software je vzhledem k jeho charakteru obtížné dostatečně vyzkoušet, nejvýznamnějším důkazem funkčnosti je především přiložená statistika získaná autorem práce. Instalační a uživatelská příručka je sice poněkud skromná, ale pro základní použití dostačující.

Text práce je z hlediska objemu mírně podprůměrný a ani objem softwaru implementovaného autorem není velký. Přesto lze konstatovat, že zadání práce bylo naplněno, neboť výsledkem je funkční systém poskytující požadované výsledky a text práce je vyhovujícím popisem problému a jeho řešení. Proto doporučuji tuto práci k obhajobě.

17.5.2010

David Bednárek