Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Marek Čermák
Název práce	Extraction and representation of unified metadata from files and file systems based on data formats
Rok odevzdání	2023
Studijní program	Informatika **Studijní obor** Softwarové a datové inženýrství

Autor posudku	Martin Svoboda **Role** Oponent
Pracoviště	KSI

Text posudku:

Cílem hodnocené diplomové práce byl návrh, implementace a evaluace rozšiřitelného nástroje, pomocí kterého by bylo možné provádět analýzy souborů nejrůznějších formátů za účelem extrakce metadat popisujících jejich skutečný obsah nebo vnitřní strukturu, konkrétně v podobě RDF trojic založených především na běžně používaných ontologiích. Tento cíl byl v plném rozsahu splněn.

Zpracování vstupních souborů je řešeno na několika vrstvách, a to od využití existujících informací na úrovni souborových systémů (např. prvotní odhady formátu souboru založené na příponách souborů), přes analýzu datového obsahu na úrovni jednotlivých bytů (detekce signatur v podobě nejrůznějších specifických sekvencí bytů apod.), až po nejrůznější specifické analýzy logického obsahu u již detekovaných formátů (na základě specifického parsování jejich obsahu).

Konkrétně je podporována řada běžně používaných formátů, např. PDF, JPG, ZIP a další. Zvláštní pozornost je pak věnována zejména formátu XML, protože je používán pro řadu odvozených formátů jako SVG, SOAP apod. Uvažuje se rekurzivní chápání obsahu souborů a jejich analýza (např. ISO soubor obsahující ZIP archiv s PDF dokumentem), stejně jako hierarchie formátů (např. JAR aplikace v Javě jakožto speciální případ ZIP archivu).

Přestože to není vždy možné, snahou je omezovat počet průchodů nutných pro zpracování jednotlivých souborů na jeden. Schopnost práce s velkými soubory potom vyplývá z použitých knihoven a jejich omezení. Kromě vlastní detekce formátů, jejich analýzy a generování zjištěných metadat v podobě RDF trojic je rovněž umožněno nad takovými grafy vyhodnocovat SPARQL dotazy. Ty jsou vyhodnocovány iterativně nad částečnými daty, ale jen naivním způsobem, tedy vždy atomicky a jen ve vztahu k omezením LIMIT nebo formě dotazů ASK. Je tedy více než diskutabilní, jestli je takové řešení přínosné.

Práce z hlediska struktury obsahuje všechny očekávané součásti. Svým rozsahem je nadprůměrná, většina textu se navíc věnuje původním myšlenkám. Pamatováno je i na testování a evaluaci celého řešení. Text práce je napsán kvalitní a plynulou angličtinou s minimem chyb (například v podobě psaní velkých písmen). Informace jsou podávány strukturovaně a dobře argumentovány. Počet citovaných zdrojů je značný, až na výjimky však jde jen o nejrůznější specifikace W3C, RFC apod.

Aplikace je navržena jako konzolová s možností ovlivňovat její chování pomocí parametrů předaných přes příkazovou řádku. K dispozici je také nadstavba umožňující používání této aplikace skrze webové rozhraní. Pro cílové uživatele bych však v takovém případě očekával kvalitnější, propracovanější a přívětivější GUI, resp. možnost ovlivňovat chování aplikace přes samostatné konfigurační soubory (kvůli pohodlnějšímu nebo opakovanému používání velkého množství parametrů).

Práce má čistě implementační charakter, neobsahuje tedy žádné zajímavé nebo netriviální aspekty z pohledu základního výzkumu. Navržená aplikace však není jen jakýmsi minimalistickým prototypem, ale představuje plnohodnotné a použitelné řešení. Z pohledu složitosti je také potřeba vyzdvihnout celkově komplexní zpracování, nutnost seznámení se s velkým množstvím nejrůznějších knihoven, stejně jako nutnost vypořádat se s řadou technických komplikací a aspektů.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Datum 27. května 2023 **Podpis**