

Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce Bc. Petr Fejfar
Název práce Interactive crawling and data extraction
Rok odevzdání 2017
Studijní program Informatika **Studijní obor** Softwarové systémy

Autor posudku Doc. Mgr. Martin Nečaský, Ph.D. **Role** Oponent
Pracoviště KSI

Text posudku:

Práce se zabývá analýzou, návrhem a implementací softwarového nástroje pro automatické vytěžování WWW stránek. Rozdílem oproti existujícím nástrojům a přístupům je zaměření se na tzv. Rich Internet Applications (RIA). Ty se vyznačují tím, že neobsahují statický nebo na serveru generovaný (a proto z pohledu klienta statický) HTML kód, ale tento kód je automaticky generován JavaScriptem na straně klienta. Současné vytěžovací nástroje se však většinou zaměřují právě na (z pohledu klienta) statický HTML kód.

Vytěžování strukturovaných dat z RIA s sebou nese řadu problémů, od nutnosti aby nástroj generoval odpovídající HTML kód, což lze vyřešit výběrem vhodné existující knihovny či nástroje, až po problém s definicí stavů RIA, ve kterých mají být strukturovaná data vytěžována a jaká konkrétní data mají být z jakého místa dynamicky generovaného HTML kódu přenášena do strukturované podoby.

Autor dostatečně analyzoval předložený problém a navrhl možná řešení. Prokázal, že je schopen se zorientovat v existující literatuře i nabídce existujících nástrojů, které může pro svoji práci využít. Vybral zajímavé řešení, které je založeno na interakci s uživatelem, čímž dokázal významně omezit stavový prostor RIA na únosnou mez. Autor v textu práce cituje a dobře sestavuje do uceleného přehledu existující literaturu z oblasti.

Práce má ale také některé nedostatky.

- 1) I když je v začátku textu práce z lingvistického pohledu v pořádku, později se úroveň anglického jazyka dosti zhorší.
- 2) Funkční i kvalitativní (nefunkční) požadavky jsou formulovány poměrně vágně. Bylo by vhodné je lépe strukturovat ve větší míře detailu, příp. použít standardizované nástroje jazyka UML.
- 3) V kapitole 4.3 jsou definovány "case studies". Čtenář by očekával, že budou využity pro evaluaci navrženého přístupu a vytvořeného softwarového nástroje. Žádná evaluace však v textu práce není.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhuje na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 24. srpna 2018

Podpis