Posudek diplomové práce
Jakub Sido: Automatická extrakce příspěvků z diskuzních fór

Ing. Ivan Habernal, Ph.D.
Technische Universität Darmstadt
Department of Computer Science
Ubiquitous Knowledge Processing Lab
Hochschulstraße 10, 64289 Darmstadt, Germany
habernal@ukp.informatik.tu-darmstadt.de

5. června 2017

Obsah práce

Cílem předložené diplomové práce bylo vyvinout a otestovat systém pro automatickou extrakci strukturovaných dat z heterogenních webových stránek. Práce v kapitolách 2–4 popisuje teoretické minimum reprezentace HTML dat a strojového učení. Kapitola 4 nabízí přehled existujících systémů a jejich kritické zhodnocení, kapitoly 5–6 pak popisují návrh nové platformy a její evaluaci.

Jako hlavní pozitivní aspekt práce hodnotím inženýrský přístup k řešení daného problému, např. využitím existujících systémů a jejich otestování na daných datech.

Zásadním nedostatkem práce je její textová složka. Celkový obsah je srozumitelný až po opakovaném čtení, na mnoha místech nemá text logickou soudržnost a návaznost, použitý styl a terminologie jsou velmi vágní.

Detailní komentáře

Abstract

Angličtina je plná chyb: "informations" – "information," "aplicated" – "applied," "intervation" – "intervention."

SOUHLASÍ
S ORIGINÁLEM

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky

Kapitola 2

V celé práci jsou věty začínající předložkou ("v", "z") malým písmenem.

Sekce 2.2 by zasloužila ustálit terminologii. Například jsem nepochopil větu "cílem získávání informací není hledat význam dat, ale pouze vyhledat atributy stejného významu."

V sekci 2.4 jsou k nalezení tyto nesmyslné věty: "Příkladem může být například tyto XPath" nebo "Navíc je přidáno několik možností navíc."

Teprve sekce 2.8 začíná konkrétním popisem, co je cílem automatické extrakce dat (čili že jde o automatcké "vytvoření" šablon pro usnadnění extrakce).

Na konci sekce 2.10 jsou zmíněny AND-OR stromy s tím, že "mají dobrou vyjadřovací schopnost, která dokáže obsáhnout vše potřebné." Čtenářovi nemusí být jasné, co je "dobrá" vyjadřovací schopnost a co je myšleno "obsáhnutím všeho potřebného;" toto je jeden příklad vágnosti, kterou zmiňuji výše.

Kapitola 3

Angličtina: "preccision" – "precision," "classifikator" – "classifier."

V sekci 3.6 je poprvé představena matice záměn se čtyřmi třídami (`NICK`, `DATE`, `COMM` a `NORM`) bez vysvětlení, co tyto třídy znamenají. Čtenář tak pouze tuší, že klasifikace probíhá právě do těchto čtyř tříd, ale co je přesně `NORM`, není nikde vysvětleno.

Kapitola 4

Sekce 4.6.1 popisuje trénovací data, ale je bohužel zmatená. Není jasné, co jsou "static pages" a "dynamic pages." V uvedeném příkladu (Obrázek 4.1) je pouze zdrojový kód s nějakými indexy.

Kapitola 5

Autor konstatuje, že v původním projektu jsou trénovací data malá a "mnohdy špatně označená" – ale chybí zdůvodnění (např. statistika).

Obrázek 5.1 obsahuje v popisu třídy `Validator` nicneříkající větu "Spouští proces validace procesu."

Oceňuji cross-domain experimenty představné v sekci 5.2.

Sekce 5.5 zdůvodňuje nutnost předzpracování stránek, ale není jasné proč – je to nutné pro RoadRunner systém? Stačí mu pouze HTML, nebo XHTML, nebo potřebuje vyrenderovaný JavaScript?

SOUHLASÍ
S ORIGINÁLEM

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky

Kapitola 6

Experiment 6.1 nedává žádný smysl. Jestliže je cílem metody extrahovat čtyři typy textu z diskusních fór, "přimíchání" Wikipedie a testování nemá žádnou logiku.

Sekce 6.3 testuje "párování" dvou výstupů aniž by bylo vůbec vysvětleno, co se tím párováním myslí. Příklad by rozhodně pomohl.

Experiment 6.7 nastavuje prahovou hranici–čeho? Je to filtrování slovníku podle unigramové pravděpodobnosti? Nebo něco jiného? Opět bohužel velmi vágní, bez jednoznačné interpretace.

Shrnutí

I přes výše zmíněné nedostatky především v prezentaci diplomové práce je vidět, že diplomant splnil zadání a vyvinul a důkladně otestoval systém pro extrakci dat z diskusních fór. Práci **doporučuji k obhajobě** a hodnotím stupněm **velmi dobře**.

Ing. Ivan Habernal, Ph.D.
Darmstadt, 5. června 2017

Otázky k obhajobě

1. Pokud byste měl dělat zadání znovu, zvolil byste opět systém RoadRunner nebo vyvinul vlastní řešení? Proč?

2. Diskusní fóra ve Vaší práci jsou s největší pravděpodobností lineární, čili se v nich opakují "boxy" s komentáři na více stránkách. Je Vaše metoda schopná zpracovávat také diskuse, které mají stromovou strukturu (např. Reddit)? Pokud ne, jak byste tento problém řešil?

SOUHLASÍ
S ORIGINÁLEM

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky