Diplomová práce - posudek vedoucího

Daniel Koukola: Porovnávání cen v Internetových obchodech

Předložená diplomová práce se zabývá problémem automatického zjišťování cen různých produktů v internetových obchodech a jejich porovnáváním. Cílem je navrhnout a otestovat různé metody strojového učení s učitelem (supervised machine learning), které by umožnily (polo-)automatickou klasifikaci. Součástí práce je také implementace prostředí pro anotaci a prezentaci získaných dat.

Autor rozdělil popis a řešení problému do dvou částí: extrakce textu ze samotných webových stránek internetových obchodů a přiřazení každého textu jednomu produktu.

Práce se skládá ze čtyř kapitol a přiloženého CD se softwarem a elektronickou verzí textu.

Po stručném úvodu do problematiky je v první kapitole rozebrán postup řešení první části problému: automatické extrakce informací o produktech z webových stránek internetových obchodů. Nejprve je navržen algoritmus pro extrakci popisků a cen z webových stránek, následuje popis použitých příznaků (rysů, features) pro algoritmy strojového učení. Dále jsou stručně popsány použité klasifikační algoritmy (support vector machines, rozhodovací stromy, naivní bayesovský klasifikátor). V závěru první kapitoly následuje vysvětlení práce Kuhn--Munkresova algoritmu pro nalezení optimálního párování v ohodnoceném bipartitním grafu. Tento algoritmus je použit pro automatické přiřazení cen k jejich popisům.

Následující kapitola popisuje řešení druhého problému: automatického přiřazení popisků k produktům, neboť stejný produkt (např. konkrétní model notebooku) může být v různých internetových obchodech označen pomocí různých popisků.

Kapitola třetí porovnává výsledky aplikace dříve popsaných algoritmů na oba problémy pomocí standardních metrik (accuracy, precision, recall, F-measure).

Čtvrtá kapitola stručně popisuje implementaci a použití vyvinutých nástrojů včetně programu pro anotaci webových stránek a prezentačního rozhraní.

Literatura obsahuje především odkazy na použité metody strojového učení.

Přiložené CD obsahuje zdrojové kódy vyvinutých programů, některé použité knihovny, ukázková anotovaná data a samotný text práce.

Hodnocení:

Práce je psána srozumitelně, některé pasáže možná až příliš stručně, nicméně pro pochopení postupu řešení dostatečně. Autor prokázal porozumění jednotlivým algoritmům strojového učení, navrhl jejich použití pro daný problém a implementoval na jejich základě komplexní řešení.

Implementace používá vhodné nástroje: Apache Tomcat server a PostgreSQL pro ukládání dat, knihovnu SWT pro anotační program a knihovny Weka a SVMLight pro algoritmy strojového učení.

Velmi pěkný je anotační nástroj, který pro zobrazení webové stránky k anotaci (vyznačení popisků a cen) používá knihovnu Gecko, což je zobrazovací jádro webového browseru. Anotátorovi je tedy prezentována stránka přesně v takové formě, jakou má k dispozici kupující v internetovém obchodě, což výrazně usnadňuje a zrychluje anotační proces.

Práci lze vytknout některé stylisticky nevhodné konstrukce vzniklé pravděpodobně při úpravě formulací (např. "Hlavním problémem spočívá v" na str. 27).

Student prokázal schopnost analyzovat zvolený problém, navrhnout a implementovat kompletní řešení a srozumitelně popsat svůj postup.

Závěr:

Předložená práce splňuje zadání i kritéria pro diplomové práce, doporučuji ji k obhajobě.

Praha, 19. 5. 2008, Miroslav Spousta, ÚFAL MFF UK