Posudek diplomové práce

Název: Porovnávání cen v internetových obchodech

Autor: Daniel Koukola

Oponent: Jiří Semecký

Matematicko-fyzikální fakulta Univerzity Karlovy, Praha

Diplomová práce se zabývá automatickým sbíráním informací o produktech v internetových obchodech a jejich porovnáváním. Práce obsahuje dvě základní části, které se uvedené problematiky týkají, jsou však navzájem nezávislé a řešitelné odděleně.

První část řeší problém automatického sbírání (crowlování) informací o produktech nabízených v internetových obchodech. Druhá část se pak zabývá algoritmickým seskupováním různých nabídek stejného zboží. Jelikož jsou tyto části na sobě nezávislé, vyjádřím se k nim odděleně.

Pro automatické sbírání informací o produktech je navržen algoritmus, který tranzitivně prochází webové stránky internetových obchodů a hledá v nich vzory, které sémanticky určují jednotlivé fragmenty textu. Toto určování je v praxi prováděno metodami strojového učení, konkrétně autor používá Support Vector Machines, rozhodovací stromy a naivní Bayesův klasifikátor. Tyto klasifikátory používají pro rozhodování seznamy příznaků, které autor získává analýzou obsahu stránek. Jelikož informace o produktech jsou hledány na stránkách se seznamem produktů, autor nakonec používá párování názvů produktů a jejich cen pomocí Kuhn-Munbkresova algoritmu.

I když je algoritmus navržen tak, aby mohl procházet různé struktury internetových obchodů (kategorizaci zboží do různých úrovní), neuvádí autor, jestli v praxi takové obecné řešení funguje. Algoritmus byl ověřen na 4 obchodech, bohužel v práci není uvedeno, o jaké obchody se jednalo, proto nelze posoudit, o jak těžký úkol se jednalo. Velký rozdíl ve výsledcích jednotlivých obchodů naznačuje, že některé obchody byly postaveny na podobném systému.

Tomuto přístupu by se dalo vytknout, že informace o produktech získává pouze ze stránek obsahujících seznamy produktů, nikoli detaily o konkrétních produktech. Takový přístup lze využít při zjednodušeném zadání, kde autor hledá pouze název produktu a cenu, ale pro účely získání detailních informací by nestačil.

Na závěr musím uvést, že ač algoritmicky zajímavý, je tento přístup v praxi těžko použitelný, protože internetové obchody ochotně poskytují informace o svých produktech v podobě strukturovaných XML souborů. To však neznamená, že by tento přístup nešel použít na podobné problémy crowlování, které přímo řešit nelze.

Druhá část práce se věnuje automatickému párování stejných položek. Autor správně popisuje problém a uvádí motivaci pro jeho vyřešení. Také ukazuje, proč nelze tento problém uspokojivě řešit klasickým vektorovým modelem a navrhuje řešení pomocí metod strojového učení.

Této problematice je v práci bohužel věnováno málo prostoru na úkor části první, což považuji za velkou škodu, protože se jistě jedná o zajímavější problém. Z mě dostupných informací tato výtka však nesměřuje ke studentovi ani vedoucímu jeho práce.

V práci jsem nenalezl informaci, jestli tuto metodu aplikoval na ručně prověřená (správná) data,

nebo na data získaná automatickou extrakcí z webových stránek a tedy obsahující zavlečené chyby.

Je škoda, že metoda řeší pouze přiřazování nalezených produktů k množině předem daných referenčních produktů a již nelze použít na obecné clusterování v případě, že referenční množinu produktů předem neznáme.

Pro experimenty autor použil ručně anotovaná data obsahující 560 položek ze 4 internetových obchodů. V popisu experimentů bohužel není vidět, s jakými údaji měl systém problémy a jak zásadní. Samotná výsledná čísla se mohou totiž výrazně lišit podle vybraných dat a granularity zboží. Veškeré položky se týkaly notebooků, kde jsou jednotlivé modely velmi těžko odlišitelné i ručně. Kdyby byla data více rozprostřena spektrem produktů, byla by úspěšnost výrazně odlišná.

Určitě by bylo by zajímavé ověřit experimenty na větších datech. V případě, že by student pokračoval v tomto výzkumu, by určitě bylo zajímavé detailně prozkoumat možnost seskupování jednotlivých produktů.

Závěr: Přes uvedené výtky se domnívám, že student prokázal schopnost samostatně zpracovat dané téma, obstarat si anotovaná data a přistoupit ke klasifikaci dat tvůrčím a zajímavým způsobem.

Jsem přesvědčen, že práce splňuje požadavky kladené na diplomovou práci na MFF UK a tvoří dobrý základ pro případné další studium. Práci rozhodně navrhuji ohodnotit známkou *výborně*.
Oponent je tvůrce serveru srovnanicen.cz.

V Praze 21. 5. 2007

RNDr. Jiří Semecký, PhD.