Posudok oponenta na diplomovú prácu:

Jan Mikšátko
Using Machine Learning Techniques to Analyze and Recognize Complex Patterns of Student E-Discussions

Obor: Teoretická informatika

Práca popisuje metodu DOCE (Detection of Clusters by Example), ktorá umožnuje nájsť klastry súvisiacich príspevkov v e-diskusiách. E-diskusie sú skupiny previazaných príspevkov niekoľkých ľudí diskutujúcich o spoločnej téme. Práca používala ako vstupy diskusie študentov získané a predspracované v rámci projektu Argunaut. V diskusiach sa hľadali sa pedagogicky významné klastry, napríklad „reťazec nesúhlasov" (Chain of opposition) a „vyjasnenie názoru po spatnej vazbe" (Clarification of opinion following feedback).

Práca sa skladá z kapitol: úvod do problematiky, popis prístupov k detekcii klastrov, popis metódy DOCE, vyhodnotenia experimentov, ďalšie možnosti pokračovania a záveru. Je písaná anglicky a prílohou je CD s programami, dátami a podrobnými výsledkami experimentov.

V úvode bola primerane vysvetlená problematika. V popise prístupov boli stručne rozobrané iné možnosti a zhrnuté, prečo bol použitý prístup pomocou (jednotlivých) príkladov klastrov. Je tam tiež uvedené, že unsupervised clustering bolo vyskúšané a neosvedčilo sa. Tretia kapitola je venovaná popisu zvolenej metódy DOCE a v nej použitým častiam, napr. grafového porovnávania, použité techniky predspracovania dát, použité miery podobnosti a prehľadávanie podgrafov založené na A* včetně heuristík. Štvrtá kapitola popisuje testovacie dáta, metodológiu experimentov a vyhodnocovania a uvádza najdoležitejšie výsledky. Ďalšie výsledky sú v prílohách na CD. Celkovo je práca veľmi medzioborová, o čom svedčí 67 položiek literatúry.

Práca bola pomerne silne naviazaná na projekt Argunaut. To jej umožnilo použiť niektoré vyvinuté nástroje, ale možno viedlo k vývoju metódy vhodnej pre tento konkrétny projekt. Inými slovami, nie je mi príliš jasné, či a jako by boli výsledky preveditelné a opakovateľné v kontexte iného projektu. Základné príčiny sú dve.
V kapitole 3 je veľká časť venovaná popisu metód grafovej podobnosti, ktorá mohla byť stručnejšia a následne venovať miesto popisu iných častí. Napr. mi nebolo jasné, jako a prečo metóda dokáže nájsť klastry aj v diskusiách s inou tématikou. Druhá výhrada sa týká dát. Vzhľadom k zašumenosti a uvedenej malej vnútornej konzistencii dát (a zhody medzi anotátormi) niekedy nebolo jasné, či sa meria kvalita dát alebo kvalita algoritmu. (Výsledky sú uvedené pre jednotlivé datasety samostatne, aj keď určované typy klastrov sú zhodné.)

Z tohoto hľadiska by prospel návrh experimentu, ktorý by DOCE porovnal s nejakým algoritmom učenia, ktorý by o jednotlivých hranách, prípadne dvojiciach vrcholov a samostatných vrcholoch, nezávisle rozhodoval, či patria do nejakého klastru. Tým by sa odlíšil výsledok získaný obecnými metódami strojového učenia a určil príspevok „grafovosti" vstupu a tohoto špecificky zameraného algoritmu. V tejto súvislosti ešte dodávam, že (ak som správne pochopil) algoritmus DOCE porovnáva vstup vždy s jedným vstupným grafom (modelom) a neumožňuje/nesnaží sa nájsť spoločné rysy roznych príkladov.

Vzhľadom k charakteru dát je DOCE (resp. QBE) asi správny prístup. Na druhej strane do značnej miery predpokladá, že výstupy bude spracovávať človek, buď anotátor

při príprave dát alebo učiteľ pri on-line použití (snaha o veľký recall aj pri malej precision).

Implementácia je v Jave a spúšťa sa dávkovo.

Konkrétne výhrady:

Pojmy klastry a detekcia klastrov sú typicky používané pre iné úlohy. Tu ide o určovanie hľadanie vzorov alebo relácií.

S 13/2.2 z textu sa zdá , že nejde o dobre definovanú úlohu.

S 16/2.3.4 Chýbal mi popis, aké atribúty sa používajú/možu používať pre vrcholy, hrany, texty v DOCE a iných prístupoch.

Okolo str. 22 dole a položky 30 sa rozsynchronizovali odkazy na literatúru.

S 41-42 Alg. A* pre prehľadávanie nie je škálovateľný pre vačšie vstupné grafy. Diplomant si tento problém uvedomil a rozobral zložitosť. Prípadne by bolo nutné použiť nejaký aproximačný algoritmus namiesto A*.

S 45/3.3.6 Rozhodnutia o parametroch sú popísané, ale nemusia byť správne. Napr. cena vypúšťania a pridávania hrán pri porovnávaní grafov može byť rozna, pretože studenti možu zabudnúť pridať hranu, ale asi ju nevypustia neúmyselne. V niektorých rysoch by šiel model zobecniť.

Doporučujem, aby práca bola prijatá jako práca diplomová.

Praha, 25.1.2008

RNDr. Jan Hric
KTIML MFF UK