Posudek bakalářské práce

Autor: Dušan Jenčík
Název: Kategorizace uživatelů s použitím posloupností stahovaných webových dokumentů
Posudek vypracoval **vedoucí práce**: Ing. Jan Šedivý, CSc

Dušan Jenčík se ve své bakalářské práci zabýval problémem kategorizace uživatelů na základě četnosti přístupů uživatelů k webovým stránkám seskupených na doménách 2. úrovně. K tomuto účelu měl k dispozici anonymizovaný dataset z reálného provozu. Student měl provést analýzu poskytnutého datasetu a zvolit vhodný postup pro rozpoznávaní témat pomocí generativních statistických metod. Jedná se o zajímavý problém demografického zařazení uživatelů na základě rozpoznaných témat.

V úvodních kapitolách se student odklání od demografického zařazení do srozumitelných kategorií a redukuje ho na úlohu obecného nalezení témat. To později zdůvodňuje nedostatečností popisu uživatelů, který by propagoval na jednotlivá témata. sekce 1.2 - “Definice problému” se zbytečně věnuje nutnosti anonymizaci dat na místo řádné definici problému, který popisuje velmi stručně.

V druhé kapitole se student věnuje převážně rešerši clickstreamu v podobě, ve které ho nemá k dispozici. Rešerši matice četnosti je věnován jeden odstavec, ve kterém popisuje co je v literatuře vysvětleno na místo kritického rozboru a shrnutí metod a jejich výsledků (pokud jsou k dispozici). Třetí kapitola celkem zbytečně rozsáhle popisuje redukci datasetu. Dále následuje popis použitých algoritmů včetně odkazu na použitou literaturu (v této sekci často citována wikipedie).

Student implementoval paralelní pLSA, čímž získal rozdělení pravděpodobnosti témat jednotlivých uživatelů. Poté témata interpretoval pomocí slov TF-IDF a párováním s databází DMOZ. Výsledky však není možné nějak rozumně posoudit, přesto že se práce měla zaobírat i problematikou volby kritérií pro posouzení témat.

Student volil nevhodnou formu popisu, některé části jsou velmi obecné nebo zbytečně rozepsané na místo jasné definice postupů často opakuje proč co ještě nemůže udělat viz např. 2. odstavec sekce 3.4.1. V některých částech opět volí nevhodný způsob popisu výsledků a pojmosloví, viz. str. 21: “Cluster 2 zobrazuje stránky zabývající se pouze vařením. Mohli bychom jásat, že přesně takové specializace hledáme. Bohužel zbylých cca 70 % clusterů se zabývají sexuálním obsahem.

Práce je členěna celkem srozumitelně, i když by bylo vhodné ji rozdělit na teoretickou a praktickou část. Např. sekce 4.3. - “Popis nalezených témat” nově uvádí problematiku sdílení stanice více uživateli, což patří spíše do sekce 1.2.1. - “Definice problému / Data”.

V práci je správně pracováno se seznamem literatury a tabulkami. Číslování kapitol a sekcí odpovídá běžným standardům. Obsahuje seznam použité literatury, na kterou se text pravidelně odkazuje.

Celkově student postupoval nesamostatně a bylo nutné s ním znovu a znovu probírat kroky řešení problému. Často bylo třeba revidovat zásadní chyby v průběžných výsledcích. Za hlavní nedostatek považuji fakt, že student nevěnoval práci dostatečné úsilí. I v závěrečných fázích práce bylo těžké se dohodnout na termínu konzultace. Se studentem jsem pracoval v průběhu prvního a druhého ročníku na návrhu aplikace pro mobilní telefony, ve které se velmi osvědčil. Možná, že volba tématu z oblasti strojového učení nebyla šťastná. Moje známka vychází především z celkového hodnocení dlouhodobého přístupu k řešení bakalářského problému. Vlastní obhajobu práce jsme zkoušeli na nečisto. Student má velmi dobré prezentační dovednosti a doufám, že je využije ve svůj prospěch. Bakalářskou práci hodnotím známkou **D - uspokojivě** a doporučuji jí k obhajobě.

Datum: Podpis: