Oponentský posudek diplomové práce

Název DP: **Recommender System for a Dating Service**

Diplomant: **Lukáš Brožovský**

Obsah práce:

Předmětem DP je návrh seznamovacího doporučovacího systému (recommender system) založeného na kolaborativním filtrování a dále experimenty se 4 reálnými datatovými kolekcemi. Autor do systému implementoval 4 základní "doporučovací" algoritmy, které odhadují neznámé ohodnocení sledovaného uživatele na základě existujících ohodnocení jiných uživatelů, kteří se jeví jako podobní (tj. ohodnocují vesměs podobně) jako sledovaný uživatel. Experimenty na datech z online seznamovacích webů ukazují na užitečnost kolaborativního filtrování a slibují využití tohoto přístupu v multi-uživatelských aplikacích, kde jsou prováděna ohodnocení výrobků/služeb/osob.

Text práce začíná úvodem do filtrování a doporučovacích systémů, kolaborativního filtrování, následuje přehled existujících systémů založených na uvedených principech. V druhé kapitole autor popisuje 4 datové sady, dvě z nich jsou charakteru vzájemného ohodnocení uživatelů v prostředí online seznamovacích webů. Kapitola 3 stručně popisuje použité algoritmy a Pearsonovu míru použitou v měření podobnosti ohodnocení dvou uživatelů. Kapitoly 4 a 5 podrobně popisují implementační detaily realizovaného systému. V kapitole 6 autor prezentuje výsledky experimentů, jednak syntetické (na základě dostupných dat), jednak empirické (na základě zapojení systému do skutečného použití v rámci online seznamovacího webu). Poslední kapitola shrnuje dosažené výsledky a diskutuje možná rozšíření.

Hodnocení:

Předmět DP je zajímavý a nový. Autor zvládnul všechny kritické součásti, tj. analýzu, implementaci a experimenty.

Autor upozorňuje na výpočetní náročnost algoritmů. Zde se musím ptát, proč na tyto náročné algoritmy byla použita Java a ne C++? Java mohla být použita pro GUI a další výpočetně nenáročný kód. Autor zmiňuje rovněž výpočetní náročnost kvůli nutnosti prohledávání matic za účelem nalezení k nejbližších sousedů. Zde bych autorovi doporučil indexační strukturu IGrid, která by dané algoritmy výrazně zrychlila, navíc je navržena pro podobnou situaci – tj. velké, řídké matice a podobná míra podobnosti.

K experimentům:

Naměřená chyba ohodnocování se pohybuje od 15% výše (předpokládám, že 100% je úplně špatné ohodnocování a 0% bezchybné – toto nebylo zmíněno). Je otázkou, zda tato chyba není stále příliš vysoká na odlišení šumu od relevantních informací. S tímto podezřením koresponduje např. autorovo zjištění, že rozdíl mezi univerzálními preferencemi všech uživatelů a jednoho konkrétního uživatele je pouhých 2,41% (strana 46). Opravdu mají všichni uživatelé stejné preference, anebo chyba > 15% prostě zprůměruje výsledky tak, jako by nezáleželo na jednotlivé uživatelově preferenci?

Další otázkou je samotná konzistence uživatelských ohodnocení. Jelikož uživatelské hodnocení osob je vysoce subjektivní, mohlo by se stát, že opakované ohodnocení osoby uživatelem by bylo odlišné v různých časových a „náladových" okamžicích. Skutečnost, že databáze neobsahovaly zdvojené ohodnocení na tomto faktu nic nemění, protože zdvojené nekonzistentní ohodnocení si můžeme představit např. jako různá ohodnocení téže osoby dvěma uživateli, přičemž zbytek jejich překrytých ohodnocení se shoduje. Nabízí se tedy otázka, do jaké míry chybu předpovědi ovlivňuje nekonzistentní chování uživatele...

Formálně je práce na velmi dobré úrovni, psaná dobrou angličtinou, pouze místy se vyskytují nevhodné gramatické obraty/termíny.

Bylo by vhodné výsledky experimentů publikovat na mezioborovém fóru informatika / psychologie / sociologie, např. konference Znalosti by mohla být ta pravá (v českém měřítku).

Podrobnější připomínky, poznámky, dotazy:

1) strana 2 nahoře – Content-based systémy jak je popisujete jsou spíše text-based systémy, kde se využívá anotací, resp. strukturované informace. Skutečné content-based systémy pracují s opravdovým obsahem, tj. v daném případě přímo s audio záznamem.
2) strana 7,8 – v případě MovieLens a Jester dat sjednocujete sémantiku uživatelů-filmů (resp. uživatelů-vtipů). Výsledkem je konverze husté obdélníkové matice na větší čtvercovou řídkou. Je opravdu vhodné takto uměle vyrábět symetrii dat? Nebylo by vhodnější ponechat původní matici? Nabízí se otázka, jaký má v této situaci význam Item-Item algoritmus...
3) sekce 2.5.3 – Ratings: je zajímavé, že sympatie uživatelů LíbímSeTi se prokázaly jako víceméně univerzální – souvisí s poznámkami výše
4) kapitola 3 je velmi stručná – stálo by zato popsané algoritmy opravdu napsat jako algoritmy
5) 3.1.1 – existuje mnoho měr podobnosti, proč byl použit pouze Pearsonův koeficient?
6) 3.4.1 – upravený Pearsonův koeficient využívá střední hodnoty pro korekci subjektivních hodnotících škál. Nestačilo by normovat uživatelské škály a použít klasický Pearsonův koeficient?
7) V kapitole 4 zcela postrádám funkční specifikaci systému nebo alespoň nějaký náznak Use Case diagramu ze softwarově inženýrského pohledu.
8) Kapitola 5 je zbytečně obsáhlá, naopak kapitola 3 je velmi stručná.

Závěr:

Práce splnila zadání a doporučuji ji k obhajobě.

V Praze dne 4. září 2006

RNDr. Tomáš Skopal, Ph.D.
oponent