Posudek na diplomovou práci

Použití metod předpovídání budoucích uživatelských hodnocení pro doporučování filmů

Bc. Martin Major

Práce se zabývá algoritmy pro předpověď budoucích hodnocení filmů a jejich praktickou použitelností v reálném prostředí. Autor přímo spolupracuje s největším tuzemským portálem (Česko-Slovenská filmová databáze, CSFD.cz), který sdružuje komunitu filmových nadšenců a zároveň buduje vlastní databázi filmů a jejich hodnocení. Experimenty byly prováděny na autentických datech zapůjčených z ČSFD.

V první kapitole autor vymezuje základní pojmy a cíle práce. Popisuje zde také důležité informace týkající se projektu ČSFD a obdobného zahraničního projektu Netflix. V kapitole druhé jsou pak uvedeny základní statistiky a informace o datech nasbíraných od uživatelů ČSFD.

Třetí kapitola se věnuje samotným algoritmům. Autor zde představuje čtyři známé algoritmy (náhodný, průměrné hodnocení, Slope One a kNN nejpodobnějších uživatelů) a vlastní algoritmus Film-DNA. Algoritmy samotné jsou popsány spíše stručně a jejich klíčové části jsou prezentovány poměrně nevhodně formou SQL dotazů. Na řadě míst by bylo vhodnější popsat algoritmus slovy případně pseudo-kódem a přesný SQL dotaz ponechat do apendixu nebo na přiložené DVD.

Vlastní algoritmus Film-DNA je popsán poměrně těžkopádně. Pro lepší přehlednost by bylo více než vhodné oddělit formální definici algoritmu od implementačních technik a optimalizací. Dále zcela postrádám podrobnější analýzu alternativ, které by objasnili, proč autor při návrhu tohoto algoritmu nezvolil jiné postupy. Jako příklad bych uvedl problém podobnosti uživatelů dle filmového vkusu. Přestože je poměrně klíčový pro předpovídání hodnocení, autor nikde nezvažuje například clustering algoritmy, které se v těchto situacích často používají.

Ve čtvrté kapitole se nachází výsledky experimentů, které měly za úkol ověřit přínos nového algoritmu. K testování byla použita data ČSFD, ze kterých bylo 1% nejnovějších hodnocení vyčleněno jako kontrolní vzorek. Kvalita výsledků byla počítána standardní metodou RMSE. Metodologii bych vytkl pouze dvě věci: Měření časové náročnosti výpočtů je potřeba opakovat více než 3x pro dosažení věrohodných čísel a umístění testovaných dat a kontrolního vzorku do jedné tabulky mohlo ovlivnit efektivitu indexačních technik použitého databázového systému.

Nově navržený algoritmus je porovnáván pouze s náhodným hodnocením a hodnocením založeným na počítání průměrného hodnocení všech uživatelů. Přestože toto porovnání má samo o sobě vypovídací hodnotu, velmi bych ocenil také porovnání s více sofistikovaným algoritmem (např. zmíněným Slope One). Výsledky porovnání také nejsou zcela uspokojivé. Nově navržený algoritmus jen velmi nepatrně překonává průměrný algoritmus a to navíc za cenu výrazně pomalejšího zpracování.

Při testování vybral autor několik konfiguračních parametrů, kterými je možné ovlivnit rychlost a přesnost nového algoritmu a empirickým prozkoumáním jejich prostoru stanovil optimální konfiguraci. Tento přístup hodnotím pozitivně, avšak zkoumaných parametrů by mělo být více, zejména když i optimální konfigurace nepřináší výrazné vylepšení výsledků oproti průměrnému algoritmu. Především zde postrádám seznam použitých parametrů pro sestavení DNA profilu uživatele a analýzu přínosu jednotlivých parametrů z hlediska zlepšení přesnosti.

Text práce je napsán v českém jazyce a to srozumitelně, bez zjevných gramatických chyb a pouze s minimálními slohovými nedostatky. Implementační část práce se skládá z návrhu databáze a SQL skriptů. Autor zde prokázal dobré praktické schopnosti při návrhu struktury databáze i při optimalizaci dotazů.

Celkově si myslím, že diplomová práce má veliký potenciál a základní myšlenka nově navrženého algoritmu je dobrá. K dokonalému výsledku ale chybí ještě poměrně dlouhá cesta a je velikou škodou, že autor nevěnoval dokončení práce více času. I přes tyto výtky se domnívám, že autor splnil zadání, a proto doporučuji práci k obhajobě.

2.1.2013

RNDr. Martin Kruliš
KSI, MFF UK