文章编号:1674-2869(2014)08-0075-04

同好推荐算法的实践

林雪云
福建师范大学福清分校,福建 福清 350300

摘 要:针对传统推荐算法在运算速度及稳定性不足等问题提出了基于矩阵模型的创新算法.通过对手机社区用户图书近一年的下载数据进行分析,依次测试每个月不同数据量下新旧算法的推荐效率,改进算法的离线计算方式,提前计量物品与物品之间的同好度表,同时,随机抽取百多名用户,计算新旧算法平均耗时表和数据量时间比指标表.实验表明,改进的算法具有明显的效率优势,不仅运算速度提高,运算结果可以重复使用,还提高了算法耗时的稳定性.算法拓展可用于商品的同好推荐,计算两物品之间的关联度,分析事件发生的影响因素等.

关键词:同好度;同好推荐算法;矩阵模型;数据挖掘;关联度
中图分类号:TP311.1 **文献标识码**:A **doi**:10.3969/j.issn.1674-2869.2014.08.014

0 引 言

当前,书目推荐已经成为图书馆书目检索系统中必不可少的栏目.甚至不止是图书馆,网络上各种各样的读书网址中都会有着书目推荐的专栏,书目推荐专栏已成为各网站争夺读者和点击率的关键所在.因为大众无论是否承认都有着一定的从众心理,这也是同好推荐备受欢迎的原因之一.

现在国内外大部分公司都有着各自的推荐算法,例如:Amazon、Netflix、lastfm、Pandora、Google,对于卓越亚马逊而言,其书目推荐技术使用的是Amazon的同好推荐技术;而Amazon被称为推荐之王,其销量的百分之三十依靠的不是别的,就是它所使用的同好推荐技术带来的,从中可以看出同好推荐算法革新的重要性[1].

通过对手机社区用户图书下载行为进行分析,然后产生相应的图书推荐,从而让用户方便的找到自己喜欢看的书.

1 同好推荐算法及问题分析

所谓同好推荐算法就是通过对用户以往行为的统计分析,利用一些数学算法预测分析出用户在未来一段时间可能的行为策略[2].当前比较流行的同好推荐算法主要分为两大方式:启发式和基于模型的方式.启发式的方法即对用户行为先进行主观预测再通过实际检验一步步接近用户最真实的状态.

而基于模型的方式则是从以往数据出发,通过对用户以往的一些行为数据的统计分析,在本文的书目同好推荐中就是对用户以往阅读书籍的分析[3],但书目同好推荐不仅仅局限于对单一用户或单一书籍的统计分析,而是把过去所有用户和所有书籍作为统计对象,于是不可避免的庞大数据源就出现了,而且这些数据相互交织,增加了对这些数据分析的难度.怎样高效稳定的得到所需要的结果就成了重中之重.

2 数据库查询计算的传统算法

产生书的推荐(喜欢这本书的人还喜欢什么书)步骤:1)获取还有哪些人看过这本书;2)获取这些人还看过哪些书;3)计算每本书对应的用户数;4)按每本书对应的用户数倒序输出.

产生用户的同好推荐:获取这个用户看过哪些书;获取看过这些书的所有用户;获取这些用户都还看过哪些书;计算每本书对应的用户数;按每本书对应的用户数倒序输出.

2.1 传统算法问题分析

传统算法优点:算法容易理解;在支持子查询的数据库容易实现.

传统算法缺点:只能在支持子查询的数据库实现,如mssql可以,mysql就不行;每计算一次书的推荐(或用户的推荐),都涉及嵌套查询,而统计数据通常都是很大(这样才准确),导致了计算速度很慢;每次查询结果不能复用.

收稿日期:2014-05-20
基金项目:福建省B类项目(JB13197)
作者简介:**林雪云**(1976-),女,福建闽侯人,副教授,硕士。研究方向:数据挖掘.

2.2 基于矩阵模型的创新算法

伴随着大数据时代的到来[4]，传统的查询算法已经远远无法满足当今世界的需求，传统的查询算法中往往伴随着子查询等等，在数据量较大的数据库中往往造成运行速度缓慢等严重问题，改变以往的子查询算法就成为重中之重.

2.2.1 *算法描述* 例如A、B、C三个用户下载过编号为101的图书，同时A用户又下载过编号为105、109的书，B用户又下载过编号为103、109的书，C用户又下载过编号为102、105、106、109的图书.那么，对A、C两用户而言，101和105这两本书的同好度为2.这里解释下同好度：两书之间的同好度就是同时读过这两本书的用户数[5].

针对编号为101的图书进行矩阵统计，矩阵图的横向和纵向都是书籍编号，按从小到大排列.横向和纵向的交叉点就是表示下载横向编号的书同时下载了纵向编号的书的用户数，即两书之间的同好度.因为同好度是相互的，所以只用了矩阵的上三角来存储.某本书的同好推荐只需要将这些数值倒序排列就可以了.

2.2.2 *程序实现* 1)这边考虑用一个字典表来保存两书之间的同好度，字典的键值是【两书籍中的小编号＋“$”＋两书籍中的大编号】，键值就是两书之间的同好度，如果用二维数组来表示矩阵有两个缺点：其一浪费存储空间，因为上面说明过矩阵的下半部分是没有数据的，其二不方便查找数据，要建立书籍ID和矩阵索引之间的关系才能定位相应的数据.

2)要产生一个书籍的推荐可以开始一个循环，开始值为所有书籍的最小编号，结束值为最大编号.

3)在循环体产生键值：【两书籍中的小编号＋“$”＋两书籍中的大编号】.

4)记录相关度.

5)倒序输出相关度，从而产生推荐.

2.2.3 *算法优点* 相比在数据库中计算，可以清楚的看到，只要计算一次两书之间的相关度就可以在后面的计算重复使用，可以大大提高计算速度.通过按图书编号从小到大来计算每本图书和其他图书之间的同好度，实验就可以只计算编号比当前编号大的图书的同好度，提高计算速度.

2.2.4 *算法拓展应用* 用户的同好推荐的实现基于图书同好推荐，本文上述算法已经可以获取每本书的相应的同好推荐，可以算出用户看过的所有书的同好推荐，产生并集，然后按同好度倒序输出[6].具体做法如下：

(1)假设实验要产生20本书作为某个用户的推荐.

(2)假设用户看过10本书.

(3)程序产生这10本书相应的同好推荐，每本书对应2本(要排除用户已经看过的书和其他书产生的推荐的书).

(4)对这20本书按同好度倒序输出.

另外，还可以用于商品(如食品、服装、电影等)的同好推荐，甚至于计算两物品之间的关联度[7]，用于分析事件发生的影响因素分析.例如：a事件的发生可能由于b或者c事件的发生，实验即可将a，b，c看做3本图书，把a事件发生且b事件发生记为1，从而得到a，b，c三者之间的同好度(即关联度)，比较关联度大小，即可知道a事件的主要影响因素[8].

3 实验评价

3.1 评价标准

评价标准1：数据量时间(平均一条数据所花费的时间)比：

$$P_1=\frac{(t_1-t_2)}{n}\Big/\frac{t_2}{n}=\frac{(t_1-t_2)}{t_1} \tag{1}$$

式(1)中n为数据量.

P_1越大表明传统算法相对于创新算法而言耗时更多.

评价标准2：耗时稳定性比：

$$P_2=\frac{\sum_{i=1}^{12}(t_{1i}-\bar{t}_1)^2}{n-1}-\frac{\sum_{i=1}^{12}(t_{2i}-\bar{t}_2)^2}{n-1} \tag{2}$$

P_2越大表明传统算法相对于创新算法而言更加不稳定[9].

其中t_{ij}中$j=1$对应传统算法，$j=2$对应创新算法，$i=1\ldots12$对应月份分为$1\ldots12$.

3.2 实验环境

实验使用手机社区用户近一年的下载数据，经过数据清洗、转换后，有效数据集1.2亿行.实验机器：操作系统Win 7，处理器为英特尔酷睿i5-2500，主频是3.3GHZ，内存为8Gb.

3.3 实验过程

将实验数据按1～12个月的顺序递增，依次测试：1个月，2个月，……，12个月不同数据量下新旧算法的推荐效率，实验输入数据如表1所示.

表1 实验数据

Table 1 Experimental data

	月份								
	1	2	3	4	5	6	7	8	9
数据量$\times10^7$	0.14	0.24	0.34	0.44	0.54	0.64	0.74	0.84	0.94

改进算法的离线计算，计算的结果在后续的

推荐中可以重复利用.

随机抽取 150 个用户,计算推荐的平均时间,具体如图 1 所示,新旧算法耗时表如表 2 所示,新旧算法 P_1 指标如表 3 所示,月数与 P_1 的关系如图 2 所示.

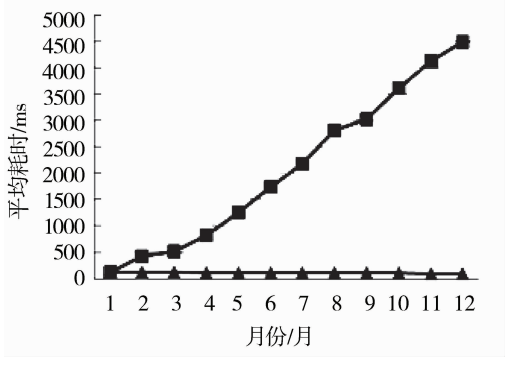

图 1　新旧算法实验分析

Fig. 1　Experimental analysis of the new and old algorithm

注:t_1　t_2

表 2　传统算法和创新算法耗时

Table 2　Time-consuming of traditional algorithms and innovative algorithms

	月　份											
	1	2	3	4	5	6	7	8	9	10	11	12
t_1	124	427	518	814	1 258	1 735	2 176	2 810	3 019	3 617	4 126	4 492
t_2	125	121	117	113	109	108	105	104	103	101	100	100
n	1 231	2 138	3 043	3 949	4 862	5 775	6 673	7 571	8 471	9 394	10 301	11 208

表 3　传统算法和创新算法 P_1 指标

Tabel 3　P_1 indicators of traditional algorithms and innovative algorithms

	月份											
	1	2	3	4	5	6	7	8	9	10	11	12
P_1	0	2.5	3.4	6.2	10.5	15.1	19.7	26	28.3	34.8	40.3	43.9

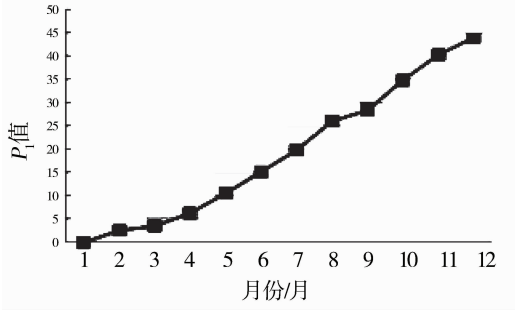

图 2　月数与 P_1 的关系

Fig. 2　Relationship of months and P_1

注:P_1

由表 3 可知,当数据量增大时,传统算法的耗时相对于创新算法是线性增长,随着数据量的继续增加,传统算法的耗时将远大于创新算法.

由公式(2)可得传统算法与创新算法稳定性对比表,即表 4 所示.

表 4　传统算法和创新算法稳定性对比

Table 4　Stability compared to traditional algorithms and innovative algorithms

	平均值	方差
传统算法	2 093	2 289 426.5
创新算法	108.8	71.2

创新算法的耗时平均值仅为传统算法的 1/20,说明创新算法的耗时远小于传统算法.同时,创新算法的方差为传统算法的 1/32135,说明创新算法的稳定性远远优于传统算法.即创新算法相对于传统算法而言更能长久保持低耗时.

3.4　实验结果

从上述实验结果可以看出:

①基于矩阵的创新算法效率明显比传统算法高.

②随着数据量的不断增大,传统算法线性下降,而基于矩阵的创新算法,由于可以重复利用计算结果,效率基本一致.

③随着数据量的不断增大,创新算法的稳定性远远优于传统算法.

4　结　语

根据大数据量下同好度推荐存在的问题,针对传统推荐算法在运算速度及稳定性不足等问题提出了基于矩阵模型的创新算法,该算法改进了传统数据库查询的推荐算法,以提高运行效率.面对的大数据,基于矩阵的创新算法,可以采用离线计算的形式,提前计算物品与物品之间的同好度表.通过实验表明,改进的算法对比传统的推荐算法具有明显的效率优势,不仅在耗时上,更在于改进算法的稳定性上.

基于矩阵模型的推荐算法不足在于在第一步新建同好度表时的耗时偏大,对这部分内容的改进也是本法未来的改进方向.

致　谢

感谢福建省自然科学基金委员会和福建师范大学福清分校科研基金的资助!

参考文献:

[1]　YOU Wen, YE Shui-sheng. A survey of collaborative filtering algorithm applied in E-Commerce recommender system [J]. Computer Technology and Development, 2006, 16(9): 70-72.

[2] HERLOCKER J L, KONSTAN J A, BORCHERS A, et al. An algorithmic framework for performing collaborative filtering [C]//Proc of the 22nd Annual Int. ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 1999:230-237.

[3] MELVILE P, MOONEY R J, NAGARAJAN R. Content-boosted collaborative filtering for lmproved recommendations [C]//Proc of the 18th National Conf on Artificial Intelligence. 2002:187-192.

[4] RESNICK P, IACOVOU N, SUCHAK M, et al. Group lens: An open architecture for collaborative filtering of net news [C]// Proc of the ACM CSCW 94 Conf on Computer-Supported Cooperative Work, New York: ACM, 1994: 175-186.

[5] GHANI R, FANO A. Building Recommender Systems Using a Knowledgebase of Product Semantics [EB/OL]. Http: www. accenture. com/ xdoc/en/services/technology/publications/recommender-ws02. pdf. 2002-10-28/2004-02-16.

[6] 李涛. 数据挖掘的应用与实践:大数据时代的案例分析[M]. 厦门:厦门大学出版社,2013.
LI Tao . Data mining application and practice: case analysis of the era of big data [M]. Xiamen: Xiamen University Press, 2013. (in Chinese)

[7] 王杨. 基于属性关联度的启发式约简算法 [J]. 计算机与数字工程, 2012(4):17-31.
WANG Yang. Heuristic reduction algorithm based on the properties of correlation [J]. Computer &Digital Engineering, 2012(4):17-31. (in Chinese)

[8] 刘臻. 计算机应用新领域-数据挖掘前景及应用探究[J]. 计算机光盘软件与应用, 2012(17): 134-136.
LIU Zhen. New areas of computer applications - data mining prospects and applications inquiry [J]. Computer CD Software and Application, 2012(17):134-136. (in Chinese)

[9] 吴昉,宋培义. 数据挖掘的应用[J]. 贵州科学,2012, 30(3):54-56.
WU Fang, Pei-yi SONG. Data mining applications [J]. Guizhou Science, 2012, 30 (3):54-56. (in Chinese)

Execution of enthusiasts recommendation algorithm

LIN Xue-yun

Fuiqing Branch of Fujian Normal University, Fuqing 350300, China

Abstract: An improved recommendation algorithm was proposed based on matrix model to improve the computing speed and stability of the traditional algorithms. With the data about the information of downloaded books for nearly a year by mobile community users, the recommendation efficiency, average time-consuming and the ratio between data size and time of the proposed algorithm were analyzed with the comparison of the traditional recommendation algorithms. The analysis results show that the recommendation efficiency is increased obviously, and the computing speed and the stability of time-consuming are also improved; moreover, the form of offline calculation is modified in the proposed algorithm, and enthusiast table between different goods is pre-calculated in offline form. The proposed algorithm can be applied in enthusiast recommendation of product, computing the correlation between the two items and analyzing the impact of factors such as events.

Keywords: enthusiasts degree; enthusiasts recommendation algorithm; matrix model; data mining; correlation

本文编辑:陈小平