基于遗传算法的聚类与协同过滤组合推荐算法

冯智明,苏一丹,覃 华,邓 海
(广西大学 计算机与电子信息学院,广西 南宁 530001)

摘 要:使用协同过滤进行推荐,在处理大数据集时存在效率问题和推荐结果质量不高的问题。k 均值聚类在处理大数据集时有着较好的性能。针对使用协同过滤进行推荐存在的问题,通过使用遗传算法将聚类和协同过滤组合起来进行项目推荐,以此来提高推荐算法的推荐效率和推荐质量,降低组合聚类和协同过滤进行推荐的复杂度。使用组合得到的算法在 MovieLens 数据集上做推荐对比实验,结果表明,相比单纯使用协同过滤进行推荐,使用基于遗传算法的聚类与协同过滤组合推荐算法进行项目推荐,能得到质量更好的推荐结果。

关键词:遗传算法;k 均值聚类;item-based 协同过滤;项目推荐
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2014)01-0035-04
doi:10.3969/j.issn.1673-629X.2014.01.009

Recommendation Algorithm of Combining Clustering with Collaborative Filtering Based on Genetic Algorithm

FENG Zhi-ming, SU Yi-dan, QIN Hua, DENG Hai
(School of Computer and Electronic Information, Guangxi University, Nanning 530001, China)

Abstract: When dealing with item recommendation with large data sets, there are problems of efficiency and the low quality of the results for collaborative filtering. K-means clustering has a better performance when processing large data sets. In order to solve problems of collaborative filtering, genetic algorithm can be used to combine clustering and collaborative filtering for item recommendation to improve the efficiency and quality of the recommendation algorithm, reduce the complexity of item recommendation by the combination of clustering and collaborative filtering. Do comparative experiments using the combination algorithm in Movielens data sets. The experimental results show that, compared with pure collaborative filtering recommendation, using genetic algorithm to combine clustering with collaborative filtering for item recommendation can get a better quality results.

Key words: genetic algorithm; k-means clustering; item-based collaborative filtering; item recommendation

0 引 言

使用传统的协同过滤进行推荐,在处理大规模数据时存在效率不高的问题,而聚类算法有着不错的分类效果且操作简单,像 k 均值聚类这样的聚类算法在处理大数据集的分类时有着较好的性能。所以,聚类算法常被用来和协同过滤组成混合的推荐算法。组合聚类和协同过滤形成的推荐算法因其较好的推荐效果和性能,在现实中有着广泛的应用,如 Google News[1]、Amazon 的推荐系统[2]等。构建一种可以简单高效地组合聚类和协同过滤进行推荐的算法有着重要意义。

1 研究背景

文献[3]提出使用 k 均值聚类来做数据平滑,以组合传统协同过滤和聚类形成推荐算法。文献[4]使用主成分分析和 k 均值聚类来组合协同过滤进行推荐。文献[5]提出使用层次聚类和语义挖掘的方法来组合协同过滤进行推荐。通过聚类和挖掘项目的隐含语义信息来提高项目相似度的计算,从而提高推荐效果。文献[6]提出使用蚁群聚类来帮助协同过滤算法挖掘用户的隐含模式,从而将相似用户分类,以提高协同过滤的推荐结果的质量。文献[7]使用聚类和免疫网络来改善使用协同过滤进行推荐时用户的邻居数目

收稿日期:2013-03-05 **修回日期**:2013-06-10 **网络出版时间**:2013-09-29
基金项目:教育部人文社会科学研究项目(11YJAZH080)
作者简介:冯智明(1987-),男,硕士研究生,研究方向为数据挖掘、数据库理论;苏一丹,教授,博士,研究方向为自然计算、电子商务;覃 华,教授,博士,研究方向为数据挖掘。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130929.1541.036.html

问题,通过保证邻居用户的多样性来改善最终协同过滤的推荐结果。文献[8]根据项目属性的相似度来对用户集合进行聚类,然后再运行协同过滤进行推荐。通过改进用户间的相似性来优化协同过滤算法的结果。文献[9]通过改进用户兴趣的多样性并使用改进的模糊聚类来搜索用户的最近邻的方法来改进协同过滤算法,使得协同过滤的推荐更适应用户兴趣的多样性,从而改善推荐结果。

使用聚类算法和协同过滤组合进行推荐,解决了协同过滤处理大规模数据时存在的效率问题,但是这种推荐方法的构建成本较高,算法本身复杂,现实中不容易使用和维护,并且这种方法的推荐结果可能会质量不高[10]。

针对上述问题,文中提出基于遗传算法的聚类与协同过滤组合推荐算法,算法通过使用遗传算法组合 item-based 协同过滤和 k 均值聚类一起进行工作,根据目标函数生成推荐结果来进行推荐。使用这种方法,聚类和协同过滤可以相对独立的运行,不需要混杂在一起,降低了推荐算法的复杂程度。而且使用遗传算法组合聚类和协同过滤进行推荐的方法所得到的推荐结果质量较高,操作过程简单。所以使用文中的方法可以解决用协同过滤进行推荐时的效率问题和推荐结果质量不高的问题,并且文中所提出的推荐算法的构建成本比较低。

2 相关背景知识

2.1 皮尔逊相关系数

为了方便程序实现,采用的是样本皮尔逊相关系数的一种简单表示:

$$r_{xy}=\frac{n\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{n\sum x_i^2-\left(\sum x_i\right)^2}\sqrt{n\sum y_i^2-\left(\sum y_i\right)^2}}$$

x_i , y_i 是向量 $\boldsymbol{x}$,$\boldsymbol{y}$ 的一个维。

文中在 k 均值聚类、item-based 协同过滤生成项目比较数据集的相似性计算都是基于样本皮尔逊相关系数。

2.2 k 均值聚类算法

给定一个观测集 $(x_1,x_2,\cdots,x_n)$,其中每个观测对象是一个 d 维向量,使用 k 均值聚类的目的就是将这 n 个元素的观测集分类成 $k(k<n)$ 个集合 $S=\{S_1,S_2,\cdots,S_k\}$,并且使得分类过程满足: $\min\limits_{S}\sum\limits_{i=1}^{k}\sum\limits_{x_j\in S_i}\|x_j-\mu_i\|^2$,其中 μ_i 是 S_i 的中心点。

2.3 item-based 协同过滤算法

item-based 协同过滤通过计算项目间的相似度来建立一个表示项目与项目间关系的比较数据集,然后使用这个项目间比较数据集结合用户的数据来推测用户的潜在兴趣。item-based 协同过滤计算加权评分的方式是 $r_{u,i}=k\sum\limits_{i'\in I}\mathrm{sim}(i,i')r_{u,i'}$,其中 $\mathrm{sim}(i,i')$ 是项目间的相似度; $r_{u,i'}$ 是用户 u 关于项目 i 的评分; $k=1/\sum\limits_{i'\in I}|\mathrm{sim}(i,i')|$,I 是项目的集合。

2.4 遗传算法

遗传算法是计算数学中用于解决最优化的搜索算法,是进化算法的一种[11]。遗传算法的操作步骤可表示为:

(1)生成初始种群个体;

(2)计算种群中个体的目标函数值;

(3)反复繁殖进化直到达到进化终止条件

①选出用于繁殖的精英个体;

②通过变异或交叉操作繁殖新种群个体;

③计算新种群个体的目标函数值;

④用新种群个体替换掉种群的非精英个体。

(4)输出最优种群个体。

3 基于遗传算法的聚类与协同过滤组合推荐算法

算法 1:使用 item-based 协同过滤生成初始种群个体。

输入:用于推荐给用户的项目集合 I。

输出:作为初始种群个体的推荐项目序列 $(x_1,x_2,\cdots,x_n)$ 。

步骤:

(1)由用户集 U,项目集 I,用户对项目的评分的集合 R 生成用户对项目的评分的关联数组 $\{u\Rightarrow\{i\Rightarrow\mathrm{rating}\}\}(u\in U,i\in I,\mathrm{rating}\in R)$;

(2)由(1)中的关联数组生成项目间的比较数据集 $\{i\Rightarrow\{i'\Rightarrow\mathrm{sim}(i,i')\}\}$,其中 $i,i'\in I$,$\mathrm{sim}(i,i')$ 是项目间的相似度;

(3)由(2)中的比较数据集生成协同过滤的推荐序列 $(x_1,x_2,\cdots,x_n)$,即遗传算法的初始种群个体。

算法 2:借助 k 均值聚类算法生成新种群个体。

输入:用于繁殖新种群个体的推荐项目序列 $(y_1,y_2,\cdots,y_n)$ 。

输出:新的推荐项目序列 $(y_1',y_2',\cdots,y_n')$,即新种群个体。

步骤:

(1)对项目集合进行 k 均值聚类,分成 k 个类 $\{S_1,S_2,\cdots,S_k\}$;

(2)根据交叉或变异需要,为序列的一个或多个元素 y_i ($i=1,\cdots,n$),在其所处的聚类 S_j ($j=1,\cdots,k$)中随机寻找用于交叉变异的新元素;

(3)交叉或变异操作,生成新的推荐序列 $(y_1', y_2', \cdots, y_n')$,即生成新的种群个体。

算法 3:基于遗传算法的 k 均值聚类与 item-based 协同过滤的组合推荐算法。

输入:用于推荐给用户的项目集合 I。

输出:经遗传算法优化后的推荐项目序列 $(z_1, z_2, \cdots, z_n)$。

步骤:

(1)由算法 1 使用 item-based 协同过滤生成初始种群个体 $(x_1, x_2, \cdots, x_n)$;

(2)计算种群中个体的目标函数值;

(3)反复繁殖进化直到达到进化终止条件

①选出用于繁殖的精英种群个体 $(y_1, y_2, \cdots, y_n)$;

②由算法 2 借助 k 均值聚类来进行交叉或变异操作繁殖新种群个体 $(y_1', y_2', \cdots, y_n')$;

③计算新种群个体的目标函数值;

④使用新种群个体替换掉种群的非精英个体。

(4)输出最优的种群个体 $(z_1, z_2, \cdots, z_n)$,即经遗传算法优化后的推荐项目序列。

4　实验及结果分析

4.1　实验目的与实验环境

该实验的目的是要证明使用遗传算法组合聚类和协同过滤算法进行项目推荐比起传统的单纯使用协同过滤进行的项目推荐,可以得到质量更好的项目推荐结果。该实验将使用单纯的协同过滤算法[12]和文中的算法作推荐对比实验,比较推荐结果的质量。

实验环境为 Windows7 平台,使用 AMD CPU Athlon II X2 250 和 2 GB 内存,程序使用 Python 编写。实验数据使用 MovieLens 100 K 数据集(包含近 1 000 名用户对近 1 700 部电影的近 10 万个评分),每位用户的评分数据的 80% 用作计算数据,20% 用作验证数据。

4.2　实验步骤

(1)确定项目推荐问题。

对比实验的项目推荐问题设计为给用户集中每位用户推荐项目集中他没看过的电影,推荐的电影的选择由目标函数决定。

问题的目标函数为收益函数(函数值越大越好),f_{obj} = avg_rating+hot,f_{obj} 是目标函数值,avg_rating 是推荐结果中电影的平均评分的平均值,hot 是推荐的电影的平均人气度。

(2)设定实验中各算法的参数。

(3)随机选出 15 位用户(从 0 计数),使用单纯的协同过滤算法为每位用户进行推荐,记录最终推荐结果的精确率和召回率。

(4)对以上选出的 15 位用户(从 0 计数),使用基于遗传算法的聚类与协同过滤组合推荐算法为每位用户进行推荐,记录最终推荐结果的精确率和召回率。

(5)使用以上结果数据绘制出统计图表。

4.3　实验结果及分析

图 1 和图 2 分别是最终推荐结果的精确率和召回率的比较(部分结果值为 0%,如用户 0)。计算精确率的公式为 precision=tp/(tp+fp),计算召回率的公式为 recall=tp/(tp+fn),其中 tp 是推荐结果中推荐正确的物品的数目;fp 是推荐结果中错误推荐的物品的数目;fn 是应该被推荐但没有出现在推荐结果中的物品的数目。推荐结果的精确率和召回率越高,则最终得到的推荐结果越准确。

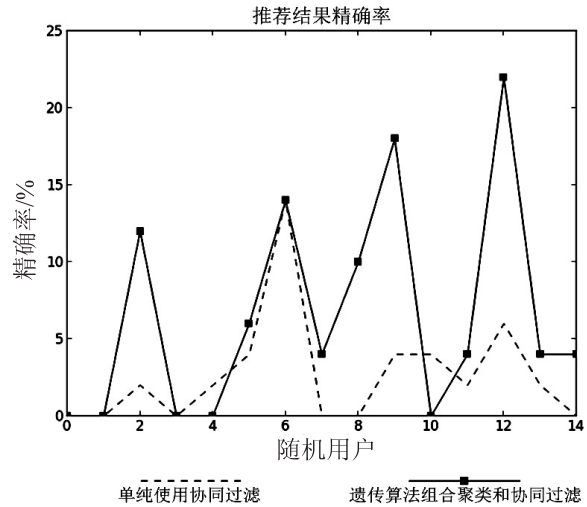

图 1　推荐结果的精确率比较

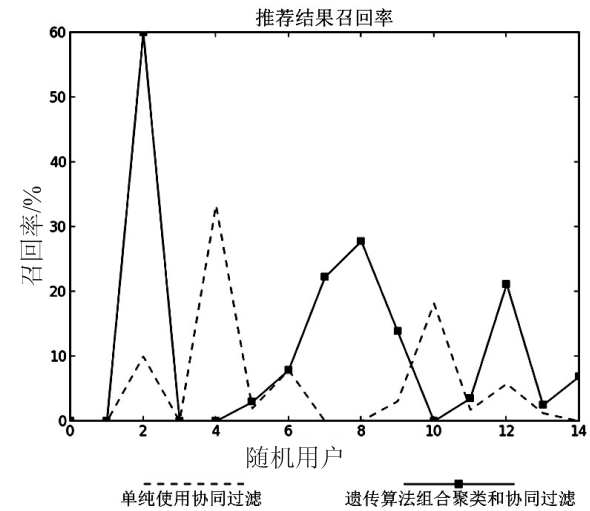

图 2　推荐结果的召回率比较

由图 1 和图 2 可以看到除了用户 4 和用户 10 外,使用由遗传算法组合聚类和协同过滤所得到的推荐算法来进行推荐,得到的推荐结果的精确率和召回率都要比单纯使用协同过滤进行推荐所得到的推荐结果的

精确率和召回率要高。所以,使用文中提出的推荐算法进行推荐,得到的推荐结果要比使用单纯的协同过滤所得到的推荐结果更准确,从而说明文中提出的推荐算法的推荐结果的质量更好。

5 结束语

针对使用协同过滤进行推荐,在处理大数据集时存在的效率问题和推荐结果质量不高的问题,文中提出基于遗传算法的聚类与协同过滤组合推荐算法来解决这些问题。实验结果表明使用文中提出的算法进行项目推荐可以得到比单纯使用协同过滤进行推荐更高质量和更准确的推荐结果,并且文中提出的算法操作简单直观,降低了组合聚类和协同过滤进行推荐所带来的高使用成本和高复杂性。

今后将继续尝试采用其他的类似遗传算法的随机优化算法来组合聚类和协同过滤进行推荐,以得到质量更高的推荐结果。

参考文献:

[1] Das A S, Datar M, Garg A, et al. Google news personalization: Scalable online collaborative filtering[C]//Proceedings of the 16th international conference on World Wide Web. New York, NY, USA: ACM, 2007: 271-280.

[2] Linden G, Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering[J]. IEEE Internet computing, 2003, 7(1): 76-80.

[3] Xue G R, Lin C, Yang Q, et al. Scalable collaborative filtering using cluster-based smoothing[C]//Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval. New York, NY, USA: ACM, 2005: 114-121.

[4] 郁 雪,李敏强. 一种结合有效降维和 K-means 聚类的协同过滤推荐模型[J]. 计算机应用研究,2009,26(10):3718-3720.

[5] 钟青燕,苏一丹,梁胜勇. 基于层次聚类和语义的标签推荐研究[J]. 微计算机信息,2010(36):199-203.

[6] 曹 波,苏一丹. 基于蚁群聚类的 top-N 推荐系统[J]. 微计算机信息,2009(9):225-226.

[7] 张 玲,王 磊,王姝媛. 基于聚类免疫网络的协同过滤推荐算法[J]. 计算机工程与应用,2008,44(27):141-144.

[8] 黄国言,李有超,高建培,等. 基于项目属性的用户聚类协同过滤推荐算法[J]. 计算机工程与设计,2010,31(5):1038-1041.

[9] 陶 俊,张 宁. 基于用户兴趣分类的协同过滤推荐算法[J]. 计算机系统应用,2011,20(5):55-59.

[10] Su X, Khoshgoftaar T M. A survey of collaborative filtering techniques[J]. Advances in artificial intelligence, 2009, 2009(4): 1-19.

[11] Whitley D. A genetic algorithm tutorial[J]. Statistics and computing, 1994, 4(2): 65-85.

[12] Segaran T. Programming collective intelligence: Building smart Web 2.0 applications[M]. Sebastopol: O'Reilly Media, Incorporated, 2007.

(上接第 34 页)

文中通过引入 DC 元数据,确保元数据的规范性,并对元数据的常见错误进行分析,在此基础上提出通过丰富系统中基础数据的建设和加强平台控制功能相结合的方式,从而达到元数据质量控制的目的。

参考文献:

[1] 孙振良. 高校机构知识库建设现状及策略研究[J]. 情报科学,2010(3):353-360.

[2] 万文娟. 学术资源开放存取的现状、障碍及策略研究[J]. 图书馆,2011(5):90-92.

[3] Jablonsk B T, Volz S, Westfechtel B. Towards a generic infrastructure for sustainable management of quality controlled primary data[C]//Proc of 3rd international workshop on ambient data integration. Crete, Greece: Springer, 2010: 130-138.

[4] 蔡迎春. 分布式机构库的质量控制[J]. 图书情报工作,2008,52(7):44-47.

[5] 陈建华. 基于 Prolog 实现语义 WEB 中的知识推理研究[D]. 北京:中国科学院,2006.

[6] 吴玲芳. 用于机构知识库的元数据研究[J]. 现代情报,2009,29(8):128-130.

[7] 庞清社. 元数据的具体功能探讨[J]. 湖北档案,2005(8):17-19.

[8] 张立肖. 基于 OAI-PMH 机构知识库互操作机制研究[J]. 河北工业科技,2010,27(2):79-82.

[9] 张毅君. 机构知识库质量评价指标研究[J]. 现代情报,2011,31(10):50-52.

[10] 郭兆红,王 欢,吕精巧. DC 元数据在数字图书馆中的应用分析[J]. 农业图书情报学刊,2009,21(9):103-105.

[11] 刘方山,孙鸿燕. DC 元数据的发展与应用[J]. 现代情报,2004(12):117-119.

[12] NISO. The framework of guidance for building good digital collections[EB/OL]. 2007-12. http://www.niso.org/publications/rp/framework3.pdf.

[13] Bruce T R, Hillmann D. The continuum of metadata quality: Defining, expressing, exploiting[M]//Metadata in practice. [s.l.]: ALA editions, 2004: 238-256.

[14] 刘家真,廖 茹. 电子文件管理元数据的质量控制与管理[J]. 图书情报知识,2009(6):91-96.

[15] 林爱群. 机构知识库元数据的自动生成与评估研究[J]. 图书馆学研究,2009(7):21-23.

基于遗传算法的聚类与协同过滤组合推荐算法

作者： 冯智明， 苏一丹， 覃华， 邓海， FENG Zhi-ming， SU Yi-dan， QIN Hua， DENG Hai
作者单位： 广西大学 计算机与电子信息学院,广西 南宁,530001

刊名： 计算机技术与发展

ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2014(1)