第46卷第1期 中国科学技术大学学报 Vol. 46, No. 1
2016年1月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jan. 2016

文章编号:0253-2778(2016)1-0082-05

基于并行化谱聚类的协同推荐算法研究

郑修猛,陈福才,黄瑞阳

(国家数字交换系统工程技术研究中心,郑州 450002)

摘要:随着大规模网络数据的增加,可扩展性成为推荐系统的一个关键因素,为此提出一种基于并行化谱聚类的协同推荐算法.首先通过并行化改进的谱聚类方法对项目进行聚类;然后在基于用户的协同推荐算法基础上,结合已聚类的项目打分信息,提出一种改进的相似用户计算方法,并进行推荐;最后在数据集上进行测试.结果表明,该算法可以有效降低时间复杂度,推荐精确度和推荐效率也有显著提高.

关键词:推荐系统;协同过滤;并行;谱聚类

中图分类号:TP18 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2016.01.011

引用格式:ZHENG Xiumeng, CHEN Fucai, HUANG Ruiyang. Research on collaborative recommendation algorithm based on parallel spectral clustering[J]. Journal of University of Science and Technology of China, 2016,46(1):82-86.

郑修猛,陈福才,黄瑞阳. 基于并行化谱聚类的协同推荐算法研究[J]. 中国科学技术大学学报,2016,46(1):82-86.

Research on collaborative recommendation algorithms based on parallel spectral clustering

ZHENG Xiumeng, CHEN Fucai, HUANG Ruiyang

(*China National Digital Switching System Engineering& Technological R&D Center, Zhengzhou* 450002, *China*)

Abstract: With the increase of large-scale network data, scalability has become a key factor in the recommendation system. A new collaborative recommendation algorithm is thus based on MapReduce parallel spectral clustering was proposed. First, items are clustered using the improved parallel spectral clustering method; Then, based on the user collaborative recommendation algorithm and combined with the clustered items' ratings, an improved calculation method for similar users is proposed to establish recommendation. The test results on the dataset show that the proposed algorithm can effectively reduce time complexity, which significantly improving its accuracy and efficiency.

Key words: recommendation system; collaborative filtering; parallel; spectral clustering

0 引言

随着Web 2.0时代的到来,网络数据呈几何指数增长,用户淹没在数据海洋中,很难获取自己真正需要的信息,造成"信息过载"现象.推荐系统通过分析用户的历史行为,将不同的项目(如电影、音乐、书

收稿日期:2015-08-27;**修回日期**:2015-09-29

基金项目:国家自然科学基金(61171108),国家重点基础研究发展(973)计划(2012CB315901,2012CB315905),国家科技支撑(863)计划(2014BAH30B01)资助.

作者简介:郑修猛,男,1989年生,硕士生. 研究方向:数据挖掘、推荐系统. E-mail:zhengxiumeng@163.com

通讯作者:陈福才,博士/研究员. E-mail:zh_xmeng@163.com

籍、新闻、图片、网页等)推荐给可能感兴趣的用户,成为解决互联网信息过载的有效途径. 协同过滤算法作为推荐系统中应用最广泛的方法之一,虽然在推荐结果上取得了较好效果,但是在精确性、扩展性、实时性等方面依然面临着一些问题. 为了解决上述问题,研究者提出了多种改进相似性的计算方法[1-4]来改善推荐的精确度. 在大数据环境下,数据更加稀疏,仅依靠评分距离很难判断两个用户之间的相似性,用户往往关注的是同一类特征的项目或者同一类特征的用户偏好更相似.

研究者将类和社区的思想引入到推荐系统中,聚类模型的协同过滤算法利用用户聚类[5]、物品聚类[6-7]或者用户和物品的联合聚类[8]来得到多个用户物品群组,提高了系统的实时性和准确度. 例如,用户聚类模型基本思想就是采用聚类算法将兴趣相似的用户聚成一类,计算目标用户和聚类后的一类相似度计算;然后根据相似偏好用户预测目标用户对该项目的评分,这就避免了相似邻居在整个用户空间的查找,提高了推荐效率. 项目聚类模型的基本思想一样,只是进行聚类的是项目而不是用户. 随着数据量的增加,传统聚类算法时间开销大,聚类结果不稳定.

谱聚类作为在 K-means 聚类算法基础上的一种改进算法,实现步骤简单,可在任意形状的样本空间上聚类,且收敛于全局最优解,非常适合用于实际问题. 由于计算和存储开销大的原因,谱聚类算法并不能直接处理大规模的数据,因此本文将谱聚类算法应用到协同推荐系统中,根据聚类结果提出了一种改进的相似用户计算方法,并在实现过程中对算法进行并行化处理.

主要工作如下:

(Ⅰ)通过项目之间相似度计算方法形成项目之间的图,利用谱聚类方法对项目进行聚类,并利用分布式思想对本文提出的谱聚类算法进行并行化处理,有效地提高了推荐的效率.

(Ⅱ)根据谱聚类处理后的结果,利用聚类簇中项目的评分信息,提出了一种新颖的计算用户之间的相似度方法,有效提高了推荐精确度.

1　相关工作

1.1　协同过滤推荐算法

协同过滤思想源于“集体智慧”. 区别于传统的基于内容的推荐算法,它是先分析用户的兴趣,寻找与目标用户偏好相似的用户;然后根据邻居用户的偏好信息,预测目标用户的潜在偏好;最后根据偏好排序产生推荐列表.

协同过滤推荐算法主要分为以下三个步骤:

(Ⅰ)用户(项目)之间的相似度计算. 相似度的计算方法主要有欧几里得距离、余弦相似性和皮尔逊相关相似性等计算方法.

(Ⅱ) 寻找最近邻居. 最近邻选择的方法可以分为:固定数量的邻居和基于相似度门槛的邻居,具体分别为选择 K 个偏好相似度靠前的用户为邻居用户或通过设定偏好相似度阈值选择大于某一阈值的用户为相似用户.

(Ⅲ) 产生推荐. 通过寻找到的最近邻用户的偏好信息,预测当前用户对其未评分项目的评分.

1.2　谱聚类算法

谱聚类的思想建立在谱图理论基础上,其本质是将聚类问题转化成一个无向图的最优划分问题[9]. 其基本思想是利用样本数据的相似矩阵(拉普拉斯矩阵)的特征向量来对数据点进行聚类.

谱聚类算法具体的实现步骤如下:

(Ⅰ)生成无向加权图. 给定一个包含 N 个点的数据集,可以把数据点看成无向加权图 $G(V,E)$,其中顶点 V 代表数据点,连接顶点之间的边为 E,E 的权重 W 代表数据点之间的相似性. 该无向图可以用邻接矩阵 $\boldsymbol{W}$ 表示,$\boldsymbol{W}$ 为 $N\times N$ 的相似性矩阵.

(Ⅱ)归一化拉普卡斯矩阵. 定义拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$,其中 $\boldsymbol{D}$ 为对角矩阵,对角元素 $\boldsymbol{D}_{ii}$ 为 $\boldsymbol{W}$ 矩阵第 i 行元素之和. 归一化矩阵拉普拉斯矩阵 $\boldsymbol{L}$ 为:

$$\bar{\boldsymbol{L}}=\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2} \tag{1}$$

(Ⅲ)计算特征值及对应特征向量. 计算归一化后的 $\bar{\boldsymbol{L}}$ 的最小 K 个特征值及其对应的特征向量 $\boldsymbol{v}_1$,$\boldsymbol{v}_2,\cdots,\boldsymbol{v}_k$,构成矩阵 $\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_k]=\boldsymbol{D}^{1/2}\boldsymbol{E}$,其中 $\boldsymbol{E}$ 为对角线为 1 的对角矩阵.

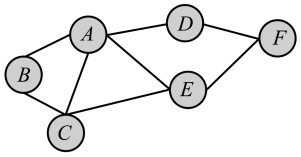
(Ⅳ)将特征向量 K-means/C-means 聚类. $\boldsymbol{V}$ 中的每一行代表数据点在 k 维空间的压缩,按行聚类,即可得到 K 个聚类结果. 通常对 $\boldsymbol{V}$ 进行归一化形成矩阵 $\boldsymbol{U}$.

$$\boldsymbol{U}_{ij}=\frac{\boldsymbol{V}_{ij}}{\sqrt{\sum_{r=1}^{k}\boldsymbol{V}_{ir}^2}},i=1,2,\cdots,n;j=1,2,\cdots,k \tag{2}$$

2 基于谱聚类的推荐模型及其并行化实现

2.1 基于谱聚类的推荐模型

给出用户-项目评分矩阵 $\boldsymbol{R}_{m\times n}$，$m$ 表示用户数目，n 表示项目数目. 对项目进行聚类分组，需要计算项目之间的相似度得到相似矩阵；进一步将项目看作图中的顶点，将项目之间的相似度看成边，便得到我们常见的图的概念，图1显示的由用户项目评分矩阵转为项目相似度图的示意.



(a) 项目图

$$\begin{bmatrix} & A & B & C & D & E & F \\ A & 0 & S_{AB} & S_{AC} & S_{AD} & S_{AE} & S_{AF} \\ B & S_{BA} & 0 & S_{BC} & S_{BD} & S_{BE} & S_{BF} \\ C & S_{CA} & S_{CB} & 0 & S_{CD} & S_{CE} & S_{CF} \\ D & S_{DA} & S_{DB} & S_{DC} & 0 & S_{DE} & S_{DF} \\ E & S_{EA} & S_{EB} & S_{EC} & S_{ED} & 0 & S_{EF} \\ F & S_{FA} & S_{FB} & S_{FC} & S_{FD} & S_{FE} & 0 \end{bmatrix}$$

(b) 邻接矩阵

图1 图模型

Fig. 1 Graph model

图1中，顶点代表项目，边代表项目之间的相似性值，邻接矩阵是对称矩阵.

在项目的相似性计算时，可以利用项目多种属性的信息，建立项目的相似性，重叠的属性越多项目越相似. 由于数据集的限制，本文利用改进的余弦相似度公式计算项目之间的相似度.

$$\mathrm{sim}(i,j)=\frac{\sum_{u\in U}(R_{u,i}-\overline{R_i})}{\sqrt{\sum_{u\in U}(R_{u,i}-\overline{R_i})^2}}\cdot\frac{(R_{u,j}-\overline{R_j})}{\sqrt{\sum_{u\in U}(R_{u,j}-\overline{R_j})^2}} \tag{3}$$

式中，U 为项目 i 和 j 都有评分的用户集合，$R_{u,i}$ 和 $R_{u,j}$ 分别表示用户 u 对 i 和 j 评分，$\overline{R_i}$ 和 $\overline{R_j}$ 表示集合 U 中所有用户对 i 和 j 的平均得分.

由此构建相似度矩阵 $\boldsymbol{W}_{m\times m}$，其中 $\boldsymbol{W}_{ij}=\mathrm{sim}(i,j)$. 由式(1)得到拉普拉斯矩阵 $\boldsymbol{L}$，由特征值问题 $\boldsymbol{LV}=\lambda\boldsymbol{V}$，求出矩阵 $\boldsymbol{L}$ 的最小的 k 个特征值和对应的特征向量 $\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_k$.

求解得到的 k 个特征值及其对应特征向量 $\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_k$ 组成矩阵 $\boldsymbol{V}[m][k]$(n 为项目编号，k 为低维特征编号)，这样就把原本 $m\times m$ 的矩阵转换为小得多的 $m\times k$ 矩阵. 本文利用 K-means 对矩阵 $\boldsymbol{V}$ 进行按行聚类，每一行相似图中的一个项目节点，其聚类结果便是谱聚类的最终输出结果. 这样便可得到项目分组.

2.2 谱聚类算法的并行化实现

传统的谱聚类算法一般仅适用于较小规模的数据集，尤其在聚类过程中存储和处理矩阵时，需要较高的空间复杂度和时间复杂度. 面对大规模数据，快速计算是一个关键问题.

本文对谱聚类协同推荐算法进行并行化改造：

(Ⅰ)在求解拉普拉斯矩阵特征值时，本文采用并行化矩阵计算工具 PARPACK 进行求解. PARPACK 可解决大规模对称、非对称和广义特征值问题. 对于标准特征问题 $Ax=\lambda x$，PARPACK 提供了两种求解方式：

①标准求解方式 $\mathrm{OP}=A$；

②位移逆求解方式 $\mathrm{OP}=(A-\delta I)^{-1}$.

(Ⅱ)在 K-means 聚类时，本文利用 MapReduce 思想[10]进行分布式改造. Map 阶段将所有待聚类的数据划分成子集合，每个 Mapper 将自己负责的子集合内的数据分别指派到附近的中心点，即打上聚类标号. Reduce 阶段收集各 Mapper 的聚类结果，根据全局聚类结果重新计算新的 K 个中心点，直到满足终止条件，停止迭代.

2.3 评分预测

针对大规模数据集，进行评分预测时，传统的基于用户的推荐系统利用用户打分之间的余弦距离作为用户之间的相似度，并不能很好地刻画用户之间的相似偏好. 本文在传统用户协同推荐算法的基础上，结合聚类结果的项目的评分，提出一种改进相似用户计算方法.

定义 2.1 ω_{uk} 表示用户 u 打分与第 k 个聚类评分的偏离程度，结果中的偏离程度，用评分误差表示.

$$\omega_{uk}=\frac{\sum_{i\in I_{uk}}(R_{ui}-\overline{R}_k)^2}{|I_u|} \tag{4}$$

式中，I_{uk} 表示用户 u 对所属第 k 个聚类结果中项目 i 的集合，$\overline{R_k}$ 是所有第 k 个聚类项目的平均得分.

定义 2.2 sim_{uv} 表示用户的特征偏好相似度，即用户 u 和 v 在对某些特征(某些种类)项目的相似偏好程度，同样用改进余弦公式表示：

$$\mathrm{sim}_{uv}=\frac{\sum_{i=1}^{k}(\omega_{uk}-\overline{\omega_u})^2(\omega_{vk}-\overline{\omega_v})^2}{\sqrt{\sum_{i=1}^{k}(\omega_{uk}-\overline{\omega_u})^2}\sqrt{\sum_{i=1}^{k}(\omega_{vk}-\overline{\omega_v})^2}} \tag{5}$$

式中，$\overline{\omega_u}$ 和 $\overline{\omega_v}$ 分别表示用户 u 和用户 v 与聚类群组偏离程度的平均值.

这样，用户 u 对 i 的预测评分最终可以表示为：

$$R_{ui}=\overline{R_u}+\frac{\sum_{v\in N_u}\mathrm{sim}_{uv}\cdot(R_{vi}-\overline{R_v})}{\sum_{v\in N_u}\omega_{uk}} \tag{6}$$

式中，$\overline{R_u}$ 表示用户 u 的平均得分，N_u 表示用户 u 的邻居用户.

算法 2.1　并行化谱聚类协同推荐算法

输入：评分矩阵 $\boldsymbol{R}$，k 为选取的最小特征值数目

输出：项目所属的 k 个聚类结果和指定用户的预测评分.

①由评分矩阵构建相似矩阵 $\boldsymbol{W}$；

②由式(1)构造拉普拉斯矩阵 $\boldsymbol{L}$；

③求解拉普拉斯矩阵的特征向量，构建特征矩阵 $\boldsymbol{V}$；利用 PARPACK 并行软件库工具，对矩阵的特征值和特征向量的分布式求解；

④基于 MapReduce 的 K-means 聚类算法，最终形成 k 个聚类结果；

⑤结合聚类结果，改进的用户相似度计算方法；

⑥评分预测.

3　实验结果及分析

3.1　实验数据

为测试本文并行化谱聚类的推荐效果和算法性能，本文分别采用 MovieLens 100 k、1 M 和 10 M 数据集，其中 100 k 中包含 943 个用户对 1 682 部电影的 10 万条评分，1 M 包含 6 040 个用户对 3 900 部电影的 100 万条评分，10 M 包含 71 567 个用户对 10 681 部电影的 1 000 万条评分. 本文随机选择 90%的数据作为训练集，10%的数据作为测试集.

3.2　评价指标

均方根误差 RMSE(Root mean squared error)是最常用的推荐质量度量方法. RMSE 越小，表明预测结果越精确. 本文采用 RMSE 作为推荐性能评价指标. 计算公式如下：

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i,j}|R_{ij}-\hat{R}_{ij}|^2}{N}} \tag{7}$$

式中，R_{ij} 是测试数据中的真实评分，$\hat{R}_{ij}$ 是通过推荐算法预测的评分，N 是测试评分数据的数目.

3.3　实验结果分析

(Ⅰ)聚类特征数目的选择

由于聚类的数目对算法的性能有重要影响，聚类数目过少，不但聚类后的某一群组中的项目数会很多，影响后续实时性，且用户相似性计算时误差会很大. 本文在 MovieLens 100 k 测试集上进行聚类数目的最优化实验. 由图 2 知，当聚类特征数目为 13 时，RMSE 最小，推荐效果最佳. 分析可知，此时聚类结果数目相对比较均匀，聚类效果最佳.

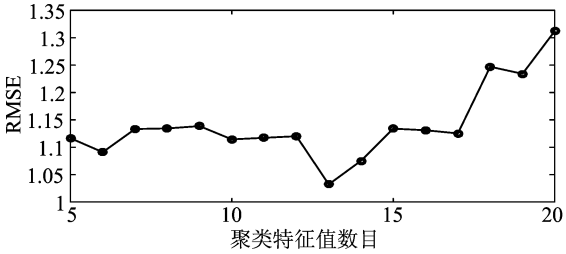
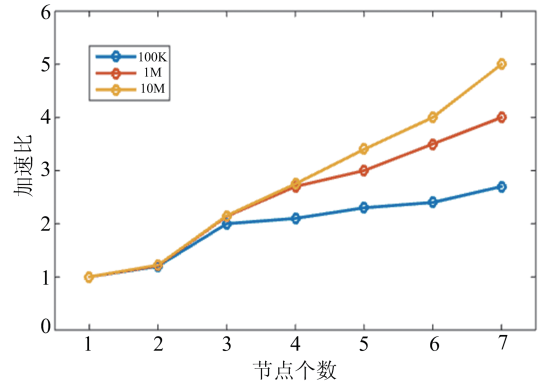


图 2　不同聚类特征数目 RMSE 的比较

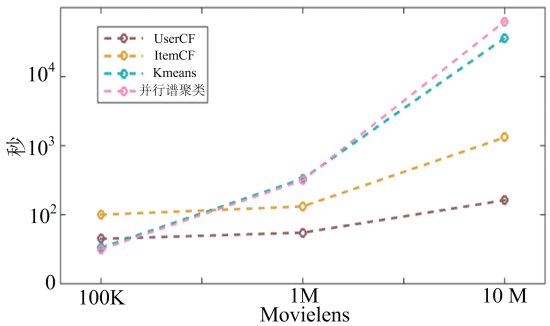
Fig. 2　The comparison of RMSE with the different number clustering features

(Ⅱ)算法效率

本文首先采用加速比 $\tau=T_1/T_N$，在三个不同数据规模下验证本文算法的可扩展性，其中 T_1 是单个节点运算所需时间，T_N 是 N 个节点的运算时间，绘制加速比曲线如图 3(a)所示. 由图可以看出，针对同一数据集，增加节点数量可提高算法效率；不同



(a) 不同规模的数据集下的加速比



(b) 不同协同推荐算法的时间比较

图 3　算法效率

Fig. 3　Algorithms efficiency

数据集,数据集规模越大,算法提升的效率越明显.本文进一步测试各算法在MovieLens 100 k、1 M、10 M三个数据集上的运行效率.从图3(b)可以看出,本文提出的算法相比传统的协同推荐算法减少了大量的时间消耗.这是由于在分布式计算时,提高了计算速度;同时,在计算用户相似度时,本文的算法复杂度由原来的$O(n^2)$降低为$O(k^2)$.

(Ⅲ)推荐算法的效果比较.

为测试本文推荐算法的推荐效果,本文与UserCF、ItemCF及K-means算法推荐效果进行对比实验,实验结果如图3所示.由图3可知,本文提出的算法相比其他算法在推荐精确度上有所提高,有效改善了推荐性能.

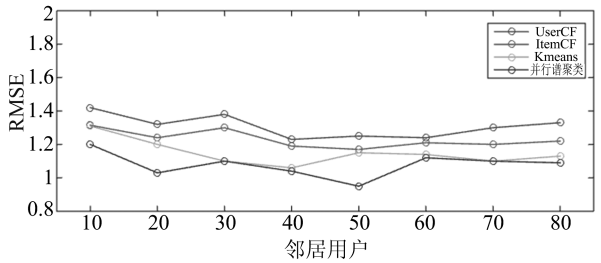


图4 不同协同推荐算法的RMSE

Fig. 4 The comparison of RMSE in different collaborative recommendation algorithms

4 结论

针对大数据背景,为解决推荐系统的扩展性问题,本文利用谱聚类方法和并行化设计思想,将谱聚类算法进行并行化处理,应用到协同推荐算法中,并通过改进用户的相似性计算方法,提高了推荐的精确度,但是在数据处理方面依然没有达到GB级甚至PB级以上更大规模的要求.下一步,我们将搭建更大规模的集群平台,尝试结合其他隐式信息进行推荐,也是进一步研究的方向.

参考文献(References)

[1] 罗辛,欧阳元新,熊璋,等. 通过相似度支持度优化基于K近邻的协同过滤算法[J]. 计算机学报,2010,33(8):1437-1445.
Luo X, Ouyang Y X, Xiong Z, et al. The effext of similarity in K-nearest-neighborhood based collaborative filtering [J]. Chinese Journal of Computers, 2010, 33(8): 1437-1445.

[2] 范波,程久军. 用户间多相似度协同过滤推荐算法[J]. 计算机科学,2012,39(1):23-26.
Fan B, Cheng J J. Collaborative filtering recommendation algorithm based on user's multi-similarity[J]. Computer Sceience, 2012, 39(1): 23-26.

[3] LiuH F, Hu Z, Mian A, et al. A new user similarity model to improve the accuracy of collaborative filtering [J]. Knowledge-Based Systems, 2014, 56(1): 156-166.

[4] 荣辉桂,火生旭,胡春华,等. 基于用户相似度的协同过滤推荐算法[J]. 通信学报,2014,35(2):16-24.
Rong H G, Huo S X, Hu C H, et al. User similarity-based collaborative filtering recommendation algorithm [J]. Journal of Communications, 2014, 35(2): 16-24.

[5] 李华,张宇,孙俊华. 基于用户模糊聚类的协同过滤推荐研究[J]. 计算机科学,2012,39(12):83-86.
Li H, Zhang Y, Sun J H. Research on collaborative filtering recommendation based on user fuzzy clustering [J]. Computer Science,2012, 39(12): 83-86.

[6] Ren X, Liu J L, Yu X, et al. ClusCite: Effective citation recommendation by information network-based clustering [C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2014: 821-830.

[7] Zhang D Q, Hsu C H, Chen M, et al. Cold-start recommendation using bi-clustering and fusion for large-scale social recommender systems [J]. IEEE Transactions on Emerging Topics in Computing, 2014, 2(2): 239-250.

[8] Gong S J. A collaborative filtering recommendation algorithm based on user clustering and item clustering [J]. Journal of Software, 2010, 5(7): 745-752.

[9] von Luxburg U. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4): 395-416.

[10] White T. Hadoop: The Definitive Guide [M]. New York: OReilly Media, 2012.