基于 MapReduce 的混合推荐算法及应用

李 程,曹 菡,师 军
(陕西师范大学 计算机科学学院,陕西 西安 710119)

摘 要:针对基于项目与基于用户两种传统协同过滤算法的不足,文中结合基于用户以及基于项目的两种传统协同过滤算法,并加以合理改进,提出了一种新型的混合型并行推荐算法。通过对新算法 MapReduce 编译,使新算法能够在 Hadoop 云平台下顺利运行。在可以利用以基于用户的方法为基础划定出定量的邻居范围,保证了推荐的个性化,同时,利用基于项目的协同过滤算法进行推荐,最终根据综合因素调整评分预测方法得出符合实际的推荐结果。实验结果表明,在数据量相对较大时新算法不仅在处理速度上表现更加优越,而且明显提高了推荐精确度。同时文中将该算法应用在西安本土旅游推荐服务上,针对西安市几大景点进行推荐,使新算法的准确性在实际应用中得到验证。

关键词:MapReduce;Hadoop;混合推荐算法;云计算
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2016)04-0074-04
doi:10.3969/j.issn.1673-629X.2016.04.016

Hybrid Recommendation Algorithm Based on MapReduce and Its Application

LI Cheng,CAO Han,SHI Jun
(School of Computer Science,Shaanxi Normal University,Xi'an 710119,China)

Abstract:For the shortcomings of traditional project-based and user-based collaborative filtering algorithm,a new parallel recommendation algorithm is proposed,combined user-based with project-based collaborative filtering algorithm and improved them. Through MapReduce compilation,the new algorithm can run in Hadoop cloud platform. To guarantee the personalized recommendation,it can take advantages of the collaborative filtering algorithm based on user defined a certain number of neighbors. At the same time,the project-based collaborative filtering algorithm is used to recommend. Finally,according to the comprehensive adjusted score prediction method,the recommended results are obtained. The experimental results show that the algorithm becomes more superior in the case of a large number of processing speed,and improves the accuracy of recommendation. Simultaneously,the algorithm is applied in local tourism of Xi'an referral service for several major attractions to recommend. The accuracy of the new algorithm has been verified in practical applications.

Key words:MapReduce;Hadoop;hybrid recommendation algorithm;cloud computing

0 引 言

随着互联网的不断迅速发展,每日产生的数据量呈指数级增长。在面对大数据的大容量、多类型天然特性时,尤其是处理 GB 级乃至 PB 级及以非结构化为主的数据时,要满足这样的高时效性变得尤为困难[1]。在稍纵即逝的市场机会和变幻莫测的大自然面前,大数据的高时效性犹如皇冠上那颗最炫耀夺目的宝石,吸引了从业者的目光。

大数据在给技术开发者带来大量丰富数据的同时,也给技术人员增加了从大数据中得到有效的用户信息与相关兴趣数据的难度[2]。将推荐系统个性化不仅能够从海量的带有很多干扰的数据中挖掘到有用信息,使得推荐具有更好的服务,也可大幅提升推荐的速度及准确度[3]。伴随着数据存储需求的不断提升,智能化商业的不断扩大,基于大数据挖掘的应用也得到了越来越广泛的研究与应用。

在云平台领域,Hadoop 是目前较为热门的研究平台[4],它以存储的廉价以及计算的高效著称。文中以大数据为背景、以云计算为手段,提出了一种新的混合推荐算法,并深入研究个性化推荐的内在原理,且在

收稿日期:2014-11-25 **修回日期**:2015-04-15 **网络出版时间**:2016-04-00
基金项目:国家自然科学基金资助项目(41271387);西安市科技计划基金资助项目(SF1228-3)
作者简介:李 程(1989-),男,硕士研究生,研究方向为高性能计算、云计算;曹 菡,博士,教授,研究方向为数据挖掘、智慧旅游、高性能计算;师 军,副教授,研究方向为智能信息处理。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160322.1517.018.html

Hadoop 云平台下设计了并行的个性化推荐算法,通过实验验证了其理论意义与实际应用价值[5]。

1　基于邻域的推荐算法

在众多的推荐算法之中,最基本的推荐算法就是基于邻域的推荐算法[6]。此类算法不仅仅是在应用领域得到了推广,而且还在研究者之间得到了较为深入的研究。此类算法包含两大类,分别为基于项目的协同过滤算法和基于用户的协同过滤算法。

1.1　基于用户的协同过滤算法

最基本的基于用户的协同过滤系统,是以兴趣相似为基础,先得到一组用户,在这个组中的用户对其命名为“邻居”。因为这些邻居是以兴趣相似划分的,所以他们之间的历史评分带有非常强的相似相关性[7]。由这些邻居之间的得分而推出的结果称之为 Top – N 推荐。为了得到更为精确的结果,可以用余弦相似法和皮尔逊相似法来测量每组用户或者邻居之间的相似度[8]。

基于用户的协同过滤算法由以下两步组成:

(1)以目标用户为中心寻找相似度高的邻居用户形成一个用户集;

(2)以这个用户集为中心向目标用户推荐用户集中目标用户没有涉及的物品或项目。

第 1 步的重点在于计算目标用户与测试用户之间的相似度。在这一步中,利用的主要是行为的相似,通过行为相似度推导出兴趣的相似度。假设有 u 和 v 两个用户,设 $N(u)$ 代表一个集合,这个集合是得到 u 用户正反馈的物品集合,设 $N(v)$ 也代表一个集合,这个集合是 v 用户正反馈的物品集合。进而通过式(1)计算出 u 和 v 的相似度:

$$w_{uv} = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|} \tag{1}$$

也可用余弦相似度计算:

$$w_{uv} = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)||N(v)|}} \tag{2}$$

在计算出用户与用户的兴趣相似度后,基于用户的协同过滤算法就默认给一个用户推荐一个与他兴趣最相似的 K 个用户喜欢的物品。式(3)度量了算法中用户 u 对物品 i 的感兴趣程度:

$$p(u,i) = \sum_{v \in S(u,K) \cap N(i)} w_{uv} r_{vi} \tag{3}$$

式中:$S(u,K)$ 为与 u 用户兴趣度较相近的 K 个用户;$N(i)$ 为与 i 物品有过打分记录的用户集合;w_{uv} 为 u 用户和 v 用户相互间的兴趣相似度;r_{vi} 代表用户 v 对物品 i 的兴趣。

尽管以用户为基础的协同过滤算法流行度很广,但同时也存在自身局限,例如可扩展性和响应性等方面[9]。为了解决这里的局限性问题,诞生出了基于项目的协同过滤算法,这种协同过滤算法就是以项目为基础建立推荐模型[10]。

1.2　基于项目的协同过滤算法

与基于用户的协同过滤算法不同,基于项目的协同过滤算法是以得分来进行推荐的,比较的是用户与用户之间对同一个项目的打分。该算法的核心是得到不同用户都打分的 K 个最相似项目[11]。基于项目的协同过滤算法采用以下两步:

(1)通过算法计算,得到项目与项目间的相似度;

(2)通过项目与项目之间的相似度再加上用户行为综合生成一个推荐列表给用户。

项目相似度定义为:

$$w_{ij} = \frac{|N(i) \cap N(j)|}{|N(i)|} \tag{4}$$

式中:分母 $|N(i)|$ 是喜欢项目 i 的用户数;分子 $|N(i) \cap N(j)|$ 是同时喜欢项目 i 和 j 的用户数。

因此式(4)可以看成是喜欢项目 i 的用户中有多少比例的用户同时也喜欢项目 j。

该式存在一定的缺陷,就是当 j 项目为热门项目时,其结果就会接近 1,因为很多人喜欢。这就会导致无论什么物品都会跟这种热门项目有着相似度较大的情况。为了不使这种情况出现,可以运用改进后的公式:

$$w_{ij} = \frac{|N(i) \cap N(j)|}{\sqrt{|N(i)||N(j)|}} \tag{5}$$

从式(5)可以看出,在此公式中对 j 项目的权重进行了惩罚,从而可以降低热门项目与其他项目相似的可能。

通过计算得出项目与项目之间的相似度后,通过式(6)得出 u 用户是否对 j 项目感兴趣:

$$p(u,j) = \sum_{i \in N(u) \cap S(j,K)} w_{ji} r_{ui} \tag{6}$$

式中:$N(u)$ 是一个物品集合,代表着用户喜欢的物品;$S(j,k)$ 是和 j 物品相似度最高的 k 个物品集合;w_{ji} 是物品 j 和 i 相互之间的相似度;r_{ui} 是 u 用户对 i 物品的感兴趣程度。

1.3　基于用户–物品的混合推荐算法

总结以上两节的分析描述,可以利用两种方法的优点,通过融合两种算法,演变出一个新的混合型协同过滤算法。这种新算法的核心就是需要计算两类推荐算法的推荐结果,进而进行结果的综合运算[12]。通过这样的综合运算,可以确保以用户与用户相似度动态计算出个性化推荐结果,同时也只要用一小部分相似用户便可以得到很好的推荐质量[13]。算法描述如下:

(1)计算目标用户与其他邻居用户的相似度;

(2)预设相似度阈值 m,若用户 b_n 的相似度大于阈值 m,则作为邻居;

(3)得到目标用户 a 的邻居数量 l;

(4)根据邻居利用算法进行预测推荐。

2 基于 MapReduce 的混合推荐算法

在分布式系统 Hadoop 运算中,第一步是初始化,将每一个 MapReduce 过程初始化为两个阶段[14],分别是一个 Map 过程和一个 Reduce 过程。其中,Map 过程实际是一个 Map 函数,Reduce 过程实际是一个 Reduce 函数。在整个 MapReduce 过程中,数据以一个<key,value>的形式进行传输,首先进入 Map 函数,经过运算再以<key_1,$value_2$>的形式导出。随后所有的<key,value>键值对经过 Hadoop,一起传输到 Reduce 阶段,经过 Reduce 函数的键值对为<key,(value list)>的形式,其结果也会以<key,value>的形式输出,形成一组一组的数据块。

在混合协同过滤算法中,两个核心步骤为基于用户-项目评分矩阵计算相似度以及基于相似度预测为评分项目的评分。这两个步骤与 MapReduce 并行处理思想是相契合的,可以编译实现。因此,在计算过程中,输入的键值对可以表示为<null,(User,Item,Scorl)>,输出的键值对可以表示为<($Item_1$,$Item_2$),Sim>。

第一次 MapReduce 得出用户对项目的评分,根据用户名进行排列;该阶段的 Map 函数将输入数据转换为相应<key,value>键值对,然后用 Reduce 函数将相同用户的项目合并,如图 1 所示。

第二次 MapReduce 得出项目间的相似度,将用户和项目的键值对转换成项目和项目之间的键值对。该阶段通过 Map 函数获得各项目之间同一用户的评分对比,随后通过 Reduce 函数得出项目之间的相似度,如图 2 所示。

最终得出用户推荐的相似列表,使用 Map 函数对目标用户评分进行预测,然后使用 Reduce 函数得出推荐结果,如图 3 所示。

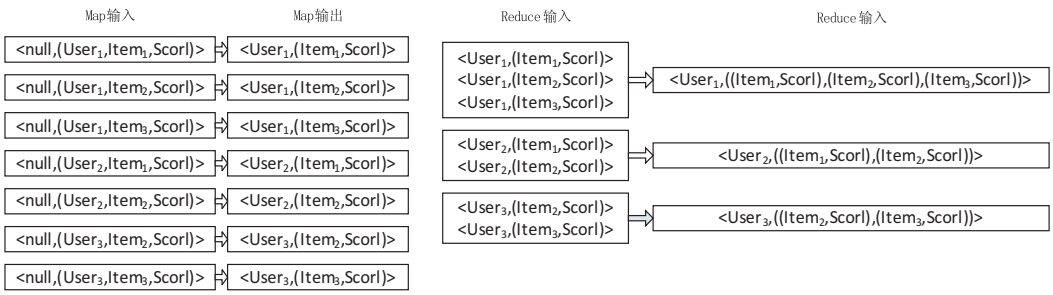

图 1 第一次 MapReduce

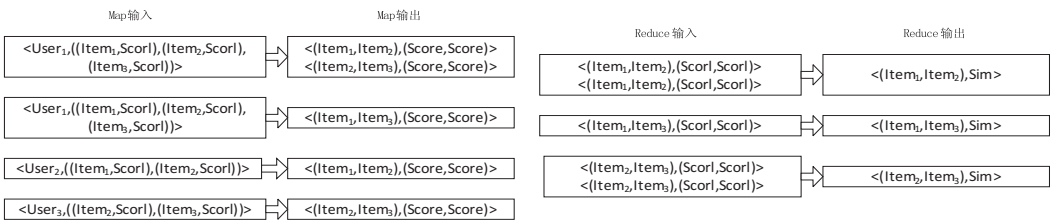

图 2 第二次 MapReduce

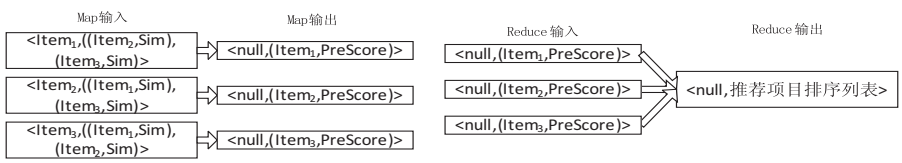

图 3 第三次 MapReduce

3 实验分析

3.1 实验设计

(1)实验准备。

一台计算机作为 NameNode 和 JobTracker 主机节点,其余九台机器作为 DateNode 和 TaskTracher 从节点。每个节点配置如下:处理器为 Intel(R) Core(TM) 2 CPU 6320 @ 1.86 GHz;内存为 1 GB;系统类型为 Ubuntu12.10,32 位操作系统。根据 Hadoop 官方网站介绍方法配置部署 Hadoop 集群版本 Hadoop1.0.2。

(2)评估标准。

为了验证实验结果的精确度,这里采用平均绝对误差(MAE)。MAE 作为推荐系统的标准评判,能够评判出推荐系统的预测精度,它的原理就是经过推导得

出预测评分与实际评分的偏差来评测算法的准确性。

$$\mathrm{MAUE}_u = \frac{\sum_{i \in \mathrm{IP}(u) \cap \mathrm{IR}(u)} |\hat{r}_{ui} - r_{ui}|}{N} \tag{7}$$

其中:IP(u) 是推荐系统为用户 u 推荐出的项目集;IR(u) 是用户 u 在测试集数据上进行评分的项目集;N 是 IP(u) 与 IR(u) 交集的项目个数。

计算出每个用户的 MAUE,然后计算该系统的 MAE:

$$\mathrm{MAE} = \sum_{u \in u} \frac{\mathrm{MAUE}_u}{|U|} \tag{8}$$

由式(8)得出:当 MAE 值越小时证明预测值与实际值差异越小。

(3)实验过程。

实验以数据的 80% 作为训练集,其余 20% 为测试集。为使实验比较更加准确,将文中算法和基于 MapReduce 用户推荐算法、基于 MapReduce 项目推荐算法,以及串行协同过滤算法一起进行实验。

3.2　实验结果

实验一对比各算法之间的 MAE 平均值,见图 4。

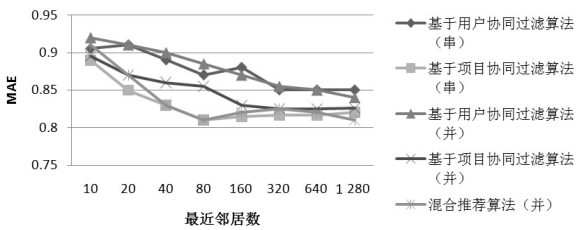

图 4　MAE 对比图

实验二在单机环境下对比各算法之间的处理时效,见图 5。

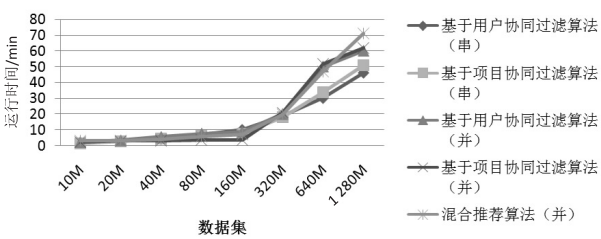

图 5　单机时效对比图

实验三在集群环境下对比各并行算法处理时效,见图 6。

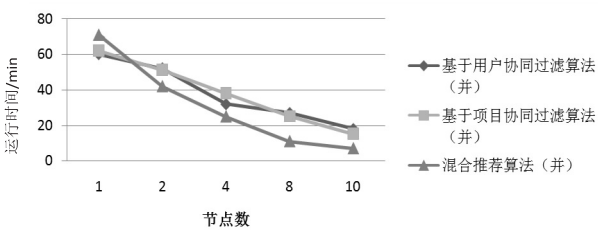

图 6　集群时效对比图

从图 4 可看出,随着邻居数的增多,新算法的推荐质量有显著提高,与其他串行算法对比推荐质量并无较大差异,并且与同为并行算法的基于 MapReduce 的项目推荐算法和用户推荐算法相比有较好表现。

从图 5 可看出,在不同数据集中随着数据量的加大,并行算法展现出其独特优势,即在较小数据集时需花费较多运算时间但在较大数据集时所需时间大大减少。

从图 6 可看出,在不同节点数数据处理对比时,证明在多节点处理上并行算法更有出色的表现。

综上得出,文中算法确实在处理大数据方面有着一定优势。

4　实例分析

最近几年,随着人民的旅游需求不断提高,旅游业蓬勃发展。“智慧旅游”以其智能、高效,得到了越来越多用户的亲睐。其中一个特色技术就是通过分析提取用户的某些资料以及以往选择,通过算法分析出一个此用户可能感兴趣的景点供其游览。在此前提下,以西安作为背景,结合在驴友网上搜集的数据,对线上的 1 785 位网友,以及在一些景点的游客,做了关于西安十五个景点的评分,过滤掉无效评分得出了 4 339 条有效数据。其中景点编号从 J_1 至 J_{15} 分别为兵马俑、华清池、回民街、半坡遗址、大雁塔、乾陵、钟鼓楼、太白山、历史博物馆、骊山、小雁塔、翠华山、秦岭野生动物园、曲江、南门,并生成评分矩阵。其中十五个景点评分为 1 ~ 5 分,不同人对不同景点心中有不同评分,其中有一些干扰数据已经排除。

从数据中可以看出,旅游者已经去过这些景点并对它们依次进行了打分,得分少的或许是没有自己游览或者只是听说。若想要为某一用户 L 推荐下一个可能喜欢的景点,使用文中算法为其推荐的是景点 12,预测评分 4.27。而实际数据其评分为 4。从结果可以看出文中算法确实能够根据旅游者的兴趣得出推荐结果,但在小规模数据上速度确实存在不足,在集群环境下速度明显低于串行算法,原因就在于 Hadoop 还需要一定时间在任务分配上,而真正的处理时间几乎微乎其微。如果要想体现出 Hadoop 处理速度的优势,数量至少要达到千万级别数据才行。所以只有在大数据量时,才可以体现出该算法的高效性以及廉价设备性。

5　结束语

针对基于项目与基于用户两种传统协同过滤算法的不足,文中提出了一种新型的混合型并行推荐算法。实验结果表明,新算法不仅在处理速度上表现更加优越,而且明显提高了推荐精确度。同时该算法在西安

(下转第 81 页)

(a)原始图像　(b)篡改图像

(c)基于传统小波变换算法的图像篡改检测结果图　(d)基于提升小波变换算法的图像篡改检测结果图

图 3　仿真结果图

复杂度低,检测结果更为准确。

参考文献:

[1] Ng T T,Chang S F,Sun Q. Blind detection of photomontage using higher order statistics[C]//Proceedings of the 2004 international symposium on circuits and systems. [s. l.]:IEEE,2004.

[2] 周琳娜. 数字图像盲取证技术研究[D]. 北京:北京邮电大学,2007.

[3] Ng T T,Chang S F. A model for image splicing[C]//Proc of international conference on image processing. [s. l.]:[s. n.],2004:24-27.

[4] 康　丽. 数字图像内容人为篡改检测[D]. 重庆:西南大学,2011.

[5] 单　薇. 基于复制粘贴的数字图像篡改检测研究[D]. 苏州:苏州大学,2014.

[6] 仝艳菲. 基于分块匹配的图像被动取证算法研究[D]. 成都:西南交通大学,2014.

[7] Christlein V,Riess C,Angelopoulou E. On rotation invariance in copy-move forgery detection[C]//Proc of IEEE international workshop on information forensics and security. [s. l.]:IEEE,2010:1-6.

[8] Al-Qershi O M,Be K. Passive detection of copy-move forgery in digital images:state-of-the-art[J]. Forensic Science International,2013,231(1):284-295.

[9] Birajdar G K,Mankar V H. Digital image forgery detection using passive techniques:a survey[J]. Digital Investigation,2013,10(3):226-245.

[10] 夏　鼐. 基于离散小波变换的图像篡改的检测[D]. 北京:中国科学院大学,2013.

[11] 高世伟,郭　雷,杜亚琴,等. 提升小波变换及其在图像处理中的应用[J]. 计算机工程与设计,2007,28(9):2066-2069.

[12] 林椹勘. 提升格式下的小波变换在图像处理中的算法研究[D]. 西安:西安电子科技大学,2005.

[13] 胡昌华,李国华,周　涛. 基于 MATLAB 7. x 的系统分析与设计—小波分析[M]. 西安:西安电子科技大学出版社,2008.

[14] Zhang T,Wang R D. Copy-Move Forgery Detection Based on SVD in digital image[C]//Proc of 2nd international congress on image and signal processing. [s. l.]:IEEE,2009:1-5.

(上接第 77 页)

本土旅游推荐服务上的应用也验证了算法的准确性。

参考文献:

[1] 陈如明. 大数据时代的挑战,价值与应对策略[J]. 移动通信,2012(17):14-15.

[2] 余肖生,孙　珊. 基于网络用户信息行为的个性化推荐模型[J]. 重庆理工大学学报:自然科学,2013,27(1):47-50.

[3] 项　亮. 推荐系统实践[M]. 北京:人民邮电出版社,2012.

[4] 刘　刚. Hadoop 应用开发技术详解[M]. 北京:机械工业出版社,2014.

[5] 陶剑文. 一种分布式智能推荐系统的设计与实现[J]. 计算机仿真,2007,24(7):296-300.

[6] 熊忠阳,刘　芹,张玉芳,等. 基于项目分类的协同过滤改进算法[J]. 计算机应用研究,2012,29(2):493-496.

[7] 黄创光,印　鉴,汪　静,等. 不确定近邻的协同过滤推荐算法[J]. 计算机学报,2010,33(8):1369-1377.

[8] Wang J,de Vries A P,Reinders M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion[C]//Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval. New York:ACM,2006:501-508.

[9] Deshpande M,Karypis G. Item-based top-n recommendation algorithms[J]. ACM Transactions on Information Systems,2004,22(1):143-177.

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

[11] Liu Q. Research on some key technologies of Chinese-English machine-in translation[D]. Beijing:Peking University,2004.

[12] Sarwar B,Karypis G,Konstan J,et al. Item-based collaborative filtering recommendation algorithms[C]//Proceedings of the 10th international worldwide web conference. [s. l.]:[s. n.],2001:285-295.

[13] 刘平峰,聂规划,陈冬林. 基于知识的电子商务智能推荐系统平台设计[J]. 计算机工程与应用,2007,43(19):199-201.

[14] 李　莉,廖建伟,欧　灵. 云计算初探[J]. 计算机应用研究,2010,27(12):4419-4422.