第 36 卷 第 3 期 华 侨 大 学 学 报（自 然 科 学 版） Vol. 36 No. 3
2015 年 5 月 Journal of Huaqiao University (Natural Science) May 2015

文章编号：1000-5013(2015)03-0292-05 **doi**：10.11830/ISSN.1000-5013.2015.03.0292

旅游大数据的 MapReduce 客户细分应用

汪永旗[1,2]，王惠娇[3]

（1. 杭州电子科技大学 自动化学院，浙江 杭州 310018；
2. 浙江旅游职业学院 旅行社管理系，浙江 杭州 311231；
3. 浙江理工大学 机械与自动控制学院，浙江 杭州 310018）

摘要： 分析 K-means 聚类算法和 Hadoop 云平台的特点，对聚类算法进行改进，给出算法的 MapReduce 实现. 通过加速比实验和旅游数据细分实验，验证了算法的有效性和高可扩展性. 针对旅游大数据的特点，构建了多指标的 RFM 扩展模型，通过文中算法聚类，得到与预期相近的聚类结果. 实验结果表明：文中算法具有较高的实用价值.

关键词： 旅游大数据；MapReduce 模型；聚类；客户细分

中图分类号： TP 39 **文献标志码**： A

在 Web 2.0 技术和移动互联网快速发展等因素的影响下，国内大型旅游 OTA 的业务量以前所未有的速度增长. 在黄金周等旅游高峰期，每天的酒店预订量可达到几十万间. 伴随着旅游消费产生了大量的过程采集、消费点评和产品推荐等数据，这些数据以各种形式保存到中心服务器上，包括文本、图片、声音、视频等. 分阶段地对这些旅游过程中产生的海量数据进行挖掘和分析是对大型线上旅游企业提出的迫切挑战[1-2]. 目前，我国大型在线旅游企业数据挖掘的数据规模已达 GB 级甚至 TB 级，传统的分析手段已难以满足现实的需要，迫切需要一种针对旅游大数据的客户细分方法，从而可以进行有效的旅游客户细分、旅游客户维护和精准营销等商业活动. 本文在应用中改进了 K-means 算法，提出了基于 MapReduce 模型的分布式聚类算法.

1 MapReduce 和 Hadoop

MapReduce 是 Google 在 2004 年的 OSDI 会议上提出的分布式并行编程模型，适用于分析处理海量数据集. MapReduce 把并行计算过程抽象为两个函数：映射（Map）和化简（Reduce）. MapReduce 就是“任务分解”模型，它通过 Map 把任务分解，用 Reduce 把处理好的结果汇总起来，得到最终结果[3-4]. 在大数据处理过程中，如果一个数据集可以分解成许多小的数据集，每个小的数据集都可以完全并行地进行处理，那么这个任务就可以用 MapReduce 来处理. MapReduce 的处理过程，如图 1 所示.

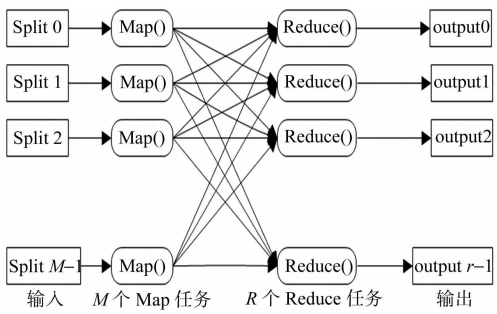

图 1 MapReduce 处理过程

Fig. 1 Processing of MapReduce

Hadoop 是 Apache 组织发布的基于 MapReduce 模型的分布式计算框架. 该架构可以在大量廉价硬件设备组成的集群上运行应用程序，为应用程序提供一组稳定可靠的接口，旨在构建一个具有高可靠性和良好扩展性的分布式系统[5]. 随着云计算的逐渐流行，这一项目被越来越多的企业所运用. Hadoop 的核心是 HDFS，MapReduce 和 HBase[6].

收稿日期： 2015-05-07

通信作者： 汪永旗(1973-)，男，讲师，博士研究生，主要从事旅游电子商务和数据挖掘的研究. E-mail：68279983@qq.com.

基金项目： 浙江省自然科学基金资助项目(Y14D060013)；浙江省教育厅校企合作项目(FW2013031)

2　聚类算法的 MapReduce 实现

K-means 算法是最经典的划分聚类算法，由于其诸多的优点，被广泛应用于客户细分等聚类应用中[7]. 因为 K-means 聚类算法具有可分解和重组的特点，所以也适合于在分布式架构下运行.

2.1　K-means 聚类算法及改进

设有 n 个对象，划分成 k 类，经过 t 次迭代，则经典 K-means 算法的时间复杂度为 $O(nkt)$. 从算法过程可以看出：算法在处理大数据集时是相对有效的，具有较好的扩展性. 计算耗时主要集中在两个环节上：一是计算各对象到中心的距离；二是将对象归类到距离最近的中心点类的过程. 对于后者，如果能减少不必要的比较和计算，则可以有效地节省时间开支. 为此，可以借用三角形三边关系定理的思想简化比较和计算过程. 具体有如下 3 个改进步骤.

步骤 1　给定含有 n 个对象的数据集 X，c_l 为 k 个初始中心，$l=1,2,\cdots,k$.

步骤 2　计算每个聚类中心的距离 $d(c_i,c_j)$，其中，$i,j=1,2,\cdots,k$.

步骤 3　计算对象 x_i 与当前所在类中心的距离 $d(x_i,c_m)$. 考察新的聚类中心 c_j，如果 $d(c_m,c_j)\geqslant 2d(x_i,c_m)$，说明 c_j 不是新的中心，可以不用计算 $d(x_i,c_j)$；否则，计算 $d(x_i,c_j)$，并与 $d(x_i,c_m)$ 比较. 继续步骤 3，直到将 x_i 归属到最近的聚类中心.

该改进算法时间复杂度为 $O(n\beta d)$. 其中：$1\leqslant\beta\leqslant k$ 是对象到中心点的计算次数. 最好的情况是计算 1 次，最坏情况下是计算 k 次，当 n 较大时，效率提高是可观的.

2.2　算法的 MapReduce 实现

用 MapReduce 处理的数据应具备以下条件：大的数据集可以被分成一个个小数据集，而且这些小数据集可以独立地被并行处理，不相互影响. 在 K-means 算法中，计算各对象到中心点的距离是被独立操作的，各对象之间没有关联[8]. 所以，K-means 算法非常适用于分布式并行计算. K-means 算法的编程思路，如图 2 所示. 由图 2 可知：在用 MapReduce 处理前，需将客户数据以行形式存储，使数据能够分片，并且各分片间数据不相关，分片过程可由 Hadoop 完成，无需另外编程.

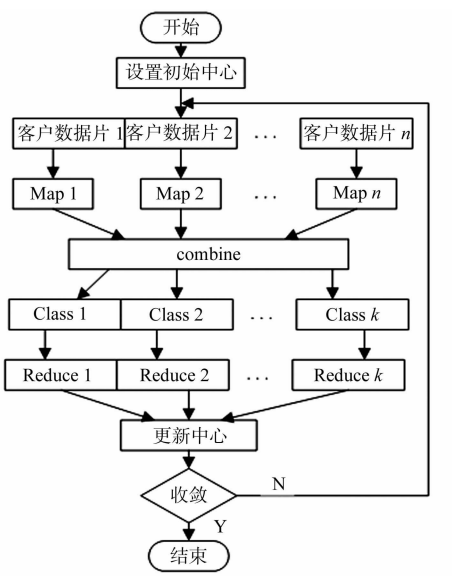

图 2　K-means 算法的 MapReduce 流程

Fig. 2　Process in MapReduce of K-means algorithm

2.2.1　*Map 函数设计*　Map 函数从特定分块中逐行读取每条记录，计算它与 k 个中心点的距离，并标明它所属的新中心类别. Map 函数的输入为原始客户数据文件和 k 个初始中心点. 原始客户数据以〈key，value〉对表示，其中：key 为记录相对于文件起始点的偏移量；value 为当前记录各维值组成的字符串. Map 函数的伪码[9]如下：

```
public void map(Writable key,Text value,Context context) {
    minDist=MAXDIST;
    for (i=0;i<k;i++) {
        if (dist(value,cluster[i])<minDist) {
            minDist= dist(value,cluster[i]);
            midClusterID= i; }
        }
    context.write(midClusterID,value);
}
```

2.2.2　*Combine 函数设计*　Combine 函数作用是对每个 Map 函数产生的结果进行本地化预处理，从而在 Reduce 时，减少不必要的通信代价，以提高整个 MapReduce 的运行性能. Reduce 函数的作用是从所有 Map 函数的结果中统计和计算出各个聚类的新中心. 为了减少通信代价，可以预先对本地 Map 函数结果进行计算，得出本地结果中各聚类对象的个数及各维数值之和，作为 Reduce 函数的输入[10-11].

Combine 函数的伪码如下

```
public void combine(Writable key,Text value,Context context) {
    num=0;
    sum: array[1..dimension];
while (value.hasnext()) {
    current= value.next();
    num++;
    for (i=0; i<dimension; i++) sum[i]+= current.value[i];
    }
context.write(key, Text(num,sum)); //输出的 value 字符串包含 num 和数组各个分量
}
```

2.2.3 Reduce 函数设计 Reduce 函数的输入是 combine 函数的输出,key 是聚簇 ID,value 中包含该簇的对象数 num 和这些对象的各维数据之和. Reduce 函数累加同一 key 的各 num 之和,并求各分量的均值,得到新的聚类中心,输出〈key,value〉对[12]. Reduce 函数的伪码为

```
public void reduce (Writable key, Text value,Context context) {
    num=0;
    while (value.hasnext()) {
        current= value.next();
        num+= current.getnum();
    for (i=0; i<dimension; i++) sum[i]+= current.value[i];
        }
    for (i=0; i<dimension; i++) mean[i]+= sum[i]/num;
    context.write(key, Text(mean));
}
```

在每次 reduce 之后,判断偏差是否小于给定的阈值. 如果小于则算法收敛;否则,把本轮 reduce 结果作为 map 的输入进行下一轮的迭代.

3 实验与分析

3.1 实验环境

文中所用实验平台是由 11 台计算机组成的千兆以太网. 其中:1 台作为 master;另外 10 台为 slaves. 各节点硬件配置:3.2 GHz Intel 双核 CPU;4 GB 内存. 软件配置:JDK 1.6.0;Hadoop 0.21.0.

实验所用的数据是 46 维的人工数据. 为了测试算法的性能,实验中构造了不同大小的数据集,包括 1,2,4,8 G. 采用加速比(speedup)作为主要的算法评价指标.

3.2 集群加速比性能实验

加速比是衡量并行系统优劣及稳定性的重要指标,是指在并行系统中,对于同一个任务,在单处理机上运行时间与在并行系统上处理时间的比率. 一方面,可以用加速比考察当系统硬件资源增加时,对相同规模任务的处理能力;另一方面,考察处理任务与硬件资源同比近似增加时,并行系统处理能力.

4 组大小成比例增长的 46 维人工数据的记录数和数据块数,如表 1 所示. 分别选择了 1,2,4,5,6 个计算节点,考量在不断增加计算节点(n)的情况下,算法的运行时间(t),得到运行时间走势图,如图 3 所示.

表 1 实验数据

Tab.1 Experimental data

编号	文件大小/GB	记录数	数据块
a	1.023	2 351 307	33
b	2.052	4 704 832	68
c	3.982	9 379 606	126
d	8.075	18 906 172	260

由图 3 可知:随着计算节点的增加,每个任务的运行时间都有显著地减少,可见 K-means 算法在 Hadoop 上运行具有较好的加速比,说

明了系统的可用性.另外,为了考察系统的扩展性,针对 a,b,c 三组数据,实验分别选择 2,4,8 个节点(n)进行运算,得到的运行时间(t),如图 4 所示.由图 4 可知:当数据规模呈正比增长时,只要相应地增加计算节点,即可保持系统的相同处理水平,体现了该 MapReduce 算法的可扩展性.

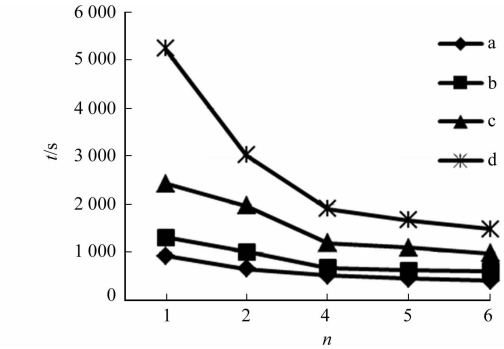

图 3　算法的运行时间走势

Fig. 3　Running time trend of the algorithm

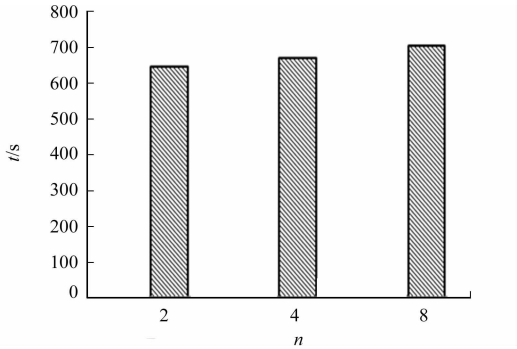

图 4　节点数与数据同比增长下算法的运行时间

Fig. 4　Running time of the algorithm in same proportion of nodes and data scale

3.3　旅游大数据客户细分实验及结果分析

实验数据来自国内某大型在线旅游网站的查询预订、过程跟踪和服务点评等数据.为了客户细分实验需要,提取了约 5 200 万条数据,涵盖了超过 120 万的客户.

首先,基于在线旅游数据的特点,在传统 RFM 模型的基础上[13-14],构建了多指标的 RFM 细分模型,如表 2 所示.进行因子分析和权重设置[15],在对初始数据进行归一化处理后,交于 Hadoop 集群处理.经过 MapReduce 算法处理后,得到 16 个客户聚类,其中的 4 个聚类在各因子上的得分和客户数(N),如表 3 所示.

表 2　多指标的 RFM 细分模型

Tab. 2　RFM model including multi index

传统 RFM	改进 RFM
R(recency)	最远消费 R_f，最近消费 R_l
F(frequency)	总体频率 Freq，月最大频率 Freq_max，月最小频率 Freq_min
M(monetary)	累计消费金额 M_sum，平均消费金额 M_avg
A(advice)	累计点评 A_sum，最近点评 A_rec

由表 3 可知:C2 类是 1 年来一直较活跃的用户,其消费额很大,频率也很高,用户较少,是公司应该重点维护的企业级客户;C5 类最近很活跃,但消费额度不大,应该是在公司点评返现推广活动(公司开展的促销活动)下,开拓的大量新进客户,这类客户的网上点评较活跃,应属于手机 APP 用户,也是企业未来发展的基石;C8 类客户曾经较活跃,有较高的消费,但最近消费很低,很可能是在今年激烈行业竞争下流失的客户;数量较大的 C11 类则属于一般价值客户.以上结果较好地反映了一年来行业的背景和企业决策所产生的影响,即在线旅游市场竞争加剧;点评返现措施带来较大业务增长;移动 APP 推广不仅吸引了大量的新客户,同时,在整个业务中的比重也有明显提高.因此,分析结果对公司新的决策有较大的参考价值.

表 3　客户聚类

Tab. 3　Customer clustering

类型	fac1	fac2	fac3	fac4	N/千人
C2	0.413	0.526	0.734	0.018	1.7
C5	0.526	0.383	0.121	0.227	131.2
C8	0.012	0.392	0.328	0.072	42.5
C11	0.138	0.176	0.189	0.180	227.2

4　结束语

利用 K-means 算法中各对象到中心点的距离是独立运算的特点,运用三边关系定理的思想改进了对象归类的过程,并给出了算法的 MapReduce 实现,通过加速比实验证明了该算法的可用性及可扩展

性.在旅游大数据客户细分应用中,构建了多指标的RFM扩展模型,经过实验,得到了预期结果.文中这种实现方法不仅可以为大型线上旅游企业提供决策支持,同时也是旅游主管部门监控、管理旅游市场的有效方法.今后将对旅游大数据挖掘中的信息安全和隐私保护问题开展研究.

参考文献:

[1] PINTO J. Analyzing Big Data is becoming a key competitive advantage[J]. Process and Control Engineering,2014,67(5):4.

[2] 孟小峰,慈祥.大数据管理:概念、技术与挑战[J].计算机研究与发展,2013,50(1):146-165.

[3] 刘鹏.实战Hadoop:开启通向云计算的捷径[M].北京:电子工业出版社,2011:60-74.

[4] LAM C. Hadoop in action[M]. Greenwich:Manning Publications Co,2011:65-72.

[5] SRIRAMA S N,JAKOVITS P,VAINIKKO E. Adapting scientific computing problems to clouds using MapReduce[J]. Future Generations Computer Systems,2012,28(1):184-192.

[6] WHITE T. Hadoop: The definitive guide[M]. Sebastopol:O'Reilly Media Inc,2012:1-39.

[7] HAN J,KAMBER M,PEI J. Data mining: Concepts and techniques[M]. Burlington:Morgan Kaufmann,2011:451-456.

[8] 江小平,李成华,向文. K-means聚类算法的MapReduce并行化实现[J].华中科技大学学报:自然科学版,2011,39(1):120-124.

[9] KHOUSSAINOVA N,BALAZINSKA M,SUCIU D. PerfXplain: Debugging MapReduce job performance[J]. PVLDB,2012,5(7):598-609.

[10] DEAN J,GHEMAWAT S. MapReduce: Simplified data processing on large clusters[J]. Communications of the ACM,2008,51(1):107-113.

[11] HUGHES,ARTHUR M. Strategic database marketing[M]. New York:McGraw-Hill Inc,2012:85-104.

[12] GUO Qi,LI Yan,LIU Tao. Correlation-based performance analysis for full-system MapReduce optimization[C]// Proceedings of IEEE International Conference on Big Data. Washington D C:IEEE Computer Society,2013:753-761.

[13] CUADROS A J,DOMINGUEZ V E. Customer segmentation model based on value generation for marketing strategies formulation[J]. Estudios Gerenciales,2014,30(130):25-30.

[14] KHOBZI H,AKHONDZADEH-NOUGHABI E. A new application of RFM clustering for guild segmentation to mine the pattern of using banks' e-payment services[J]. Journal of Global Marketing,2014,27(3):178-190.

[15] KLAS H,BJIRN L,DAG E,et al. Customer segmentation based on buying and returning behaviour[J]. International Journal of Physical Distribution and Logistics Management,2013,42(10):852-865.

Application of Tourist Segmentation Based on MapReduce under Big Data of Tourism

WANG Yong-qi[1,2], Wang Hui-jiao[3]

(1. School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China;
2. Department of Travel Agency Management, Tourism College of Zhejiang, Hangzhou 311231, China;
3. School of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: First, the characteristic of K-means clustering algorithm and Hadoop cloud platform is analyzed in this paper, the improvement of K-means clustering algorithm and its implementation of MapReduce are given. Then, the experiments of speedup and tourist segmentation are given to illustrate the effectiveness and the high scalability of the proposed method. Finally, according to the characteristics of tourism big data, a multi index RFM model is built, the clustering results which are expected indicate that the algorithm is highly practical.

Keywords: tourism big data; MapReduce model; clustering; customer segmentation

(**责任编辑**:黄晓楠 **英文审校**:吴逢铁)