第49卷第2期 中国科学技术大学学报 Vol. 49, No. 2
2019年2月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Feb. 2019

文章编号:0253-2778(2019)02-0112-07

基于 MapReduce 的商品评论热点发现算法研究

苏 浩,刘其成,牟春晓

(烟台大学计算机与控制工程学院,山东烟台 264005)

摘要:提出一种基于 MapReduce 框架的商品评论热点发现并行算法——PR-HD 算法. PR-HD 算法使用爬虫技术提取某电商平台下某热门手机的评论数据生成评论数据集,以 TF-IDF 算法来计算特征词的权重,通过特征词添加位置权重的方式来得到特征词的最终权值,建立向量空间模型(VSM)计算不同评论语句的相似度,使用 Canopy 算法和 K-means 算法相结合从而实现商品评论的热点发现. 这使得产品开发人员可以从中获取更直接有效的建议和反馈.

关键词:评论热点发现;MapReduce;Canopy 算法;K-means 算法

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2019.02.005

引用格式:苏浩,刘其成,牟春晓. 基于 MapReduce 的商品评论热点发现算法研究[J]. 中国科学技术大学学报,2019,49(2):112-118.

SU Hao, LIU Qicheng, MU Chunxiao. Research on product reviews hot spot discovery algorithm based on MapReduce[J]. Journal of University of Science and Technology of China, 2019, 49(2): 112-118.

Research on product reviews hot spot discovery algorithm based on MapReduce

SU Hao, LIU Qicheng, MU Chunxiao

(*School of Computer and Control Engineering, Yantai University, Yantai* 264000, *China*)

Abstract: A parallel algorithm based on MapReduce framework for finding hot spots from commodity reviews (PR-HD algorithm) is proposed. The PR-HD algorithm uses crawler technology to extract an electricity supplier. A review data set is generated from the review data of a popular mobile phone under the platform, and the weight of the feature words is calculated by the TF-IDF algorithm. The final weights of the feature words are obtained by adding position weights of the feature words, and a vector space model (VSM) calculation is established. The similarity of different comment sentences is combined using Canopy algorithm and K-means algorithm to realize hot spot discovery from commodity reviews. This allows product developers to obtain more direct and effective suggestions and feedback.

Key words: comment hot spot found; MapReduce; Canopy algorithm; K-means algorithm

收稿日期:2018-06-15;**修回日期**:2018-09-18

基金项目:山东省自然科学基金(ZR2016FM42);山东省重点研发计划(2016GGX109004);国家海洋局"十三五"海洋经济创新发展示范重点项目(YHC-ZB-P201701);国家自然科学基金(61702439)资助.

作者简介:苏浩,男,1995 年生,硕士生,研究方向:并行计算与数据挖掘. E-mail:1406205897@qq.com

通讯作者:刘其成,博士/教授. E-mail:ytliuqc@163.com

0　引言

近年来，随着信息技术的高速发展，互联网技术也逐渐普及、完善，电子商务逐渐发展壮大. 现如今，网络购物的方便和快捷等优点与现代人群更忙碌和紧凑的生活特征使得电子商务的需求越来越高，这给各大电商平台的发展提供了很好的机会，但同时也伴随着激烈的竞争. 为了能够在这样激烈的战争中占得先机，各平台和产品供应商大力提高自己的产口和服务质量，其中最直接有效方法之一就是从评价中获知消费者的需求和心声.

快速准确地分析现有数据，最大限度地获取其中蕴含的价值，已经成为许多公司和学者共同面对的问题[1]. 打开当前任意一个热门购物网站，可以发现大部分的商品之下都有着数量庞大的评论，一些极其热门的产品下的评论数量甚至可以达到几十万，如此庞大的数据量让商品开发者难以及时有效地获取到商品反馈信息. 这些大规模数据中，往往隐藏着商业价值重大的知识和内涵. 商品评论热点发现算法，可以有效地进行信息过滤，解决仅依靠人力所不可能完成的评论分析. 商品评论热点发现算法将海量评论信息预处理后进行文本聚类，再从每一不同类别中挖掘出评论重点，从而让商品开发者快速了解用户需求.

本文提出了基于 MapReduce 的商品评论热点发现算法——PR-HD（product review hot spot discovery）算法. 该算法将文本聚类算法与 MapReduce 分布式计算框架相结合，通过多台计算机对商品评论数据集进行深入挖掘，从而实现商品评论的热点发现. 这有利于商品开发者及时获取产品的需求与建议.

1　相关工作

在第 5 次信息技术革命、全球信息化浪潮和中国国家信息化战略的推动下，中国电子商务的理论研究和实践运作随之兴起[2]，商品评论的挖掘成为备受关注的一个研究热点. 商品评论的挖掘就是将互联网上采集用户评论作为挖掘对象，从大量评论信息中发现有关产品各方面功能和性能的评价信息的过程[3].

近年来，商品评论的价值挖掘所衍生出的算法层出不穷. 文献[4-5]利用情感词典对商品评论文本进行处理和表示，通过赋予情感词典中不同情感词的权重并结合贝叶斯分类模型对评论文本进行情感分类，将商品评论划分为积极评论与消极评论两个方面. 文献[6-7]选取目前在情感分析领域广泛应用的 SVM 作为基学习器，进一步提高了网络用户评论情感分析的准确度. 由于商品评论往往涉及产品的各个方面，以上文献只是将评论根据其情感划分为两个方面，对评论的挖掘不够全面.

为了能够更充分地挖掘评论数据隐藏的数据价值，文献[8]提出使用 K-means 聚类算法来进行商品评论的挖掘. 通过 K-means 算法对评论数据进行聚类，从而增加了商品评论挖掘的多样性. 随着数据量的增大，算法的准确率面临着新的挑战，文献[9-10]采用改进 K-means 算法的初始中心点选取方法改进算法，得到更好的聚类效果，在商场数据分析上可以进一步提高准确率. 文献[11]提出使用 Canopy 算法快速迭代出聚类中心，实现数据集的粗略聚类. 虽然准确率不尽人意，但是其在自动获取聚类中心上面有着一定的优越性. 文献[12]在分析同样分析 K-means 聚类的适用性及优劣势的基础上，提出了一种基于层次化 AP 的聚类模型，此聚类模型的第一层为 K-means 聚类模型，第二层为 AP 聚类模型，最后是聚类结果的回溯及标签化. 以上工作大大提高了评论挖掘的准确率，但是在大数据环境下仍然存在着数据规模的局限性以及难以人为确定聚类个数等问题.

本文提出了基于 MapReduce 的商品评论热点发现算法，目的是并行地对商品评论进行多方面的深入价值挖掘. 首先，PR-HD 算法可以从不同方面挖掘商品评论的隐藏价值，将商品评论的挖掘由单方面的情感分析扩展到多方面的热点发现上来，从而解决评论挖掘不够充分的问题. 其次，PR-HD 算法基于 MapReduce 框架设计，在处理超大规模的数据时展现了良好的性能，解决了串行算法无法满足大规模数据处理的问题. 同时，PR-HD 算法综合使用了 Canopy 算法来确定商品评论的热点个数，从而解决了大规模商品评论数据无法确定评论热点个数的问题，因此 PR-HD 算法更符合商品评论挖掘算法的现实环境需求.

2　相关知识

2.1　文本向量化

向量空间模型（vector space model，VSM）是一种常用的文本向量化的方法. 它将所有文本内容看

作多个单词的集合，每一个单词都分配给一个单独的索引值，这个索引值指向这个单词所对应的向量维度，一篇文章中所有单词的维度构成这篇文章的向量. 对于向量中每一个单词，我们采用词频-逆文档频率（term frequency-inverse document frequency，TF-IDF）算法来计算它所对应的维度上的值.

假设一共有 N 个文本，这些文本中共有 i 个单词，分别记为 $w_1, w_2, w_3, \cdots, w_i$ 它们的频率分别为 $f_1, f_2, f_3, \cdots, f_i$. 对每一个单词 w_i 来说，其所对应的维度上的值 W_i 可以由公式 1 求得：

$$W_i = \mathrm{TF}_i \cdot \mathrm{IDF}_i = \mathrm{f}_i \cdot \log \frac{N}{\mathrm{DF}_i} \tag{1}$$

式中，TF_i 为单词 w_i 的词频，相对应的值为单词的频率 f_i. DF_i 是单词 w_i 的文档频率，指的是在 N 个文档中包含此单词的文档个数. IDF_i 为单词 w_i 所对应的逆文档频率，我们使用 $\log \frac{N}{\mathrm{DF}_i}$ 来表示.

2.2 Canopy 算法

Canopy 算法具有速度快，易实现等优点，是一种实用的聚类算法. Canopy 算法需要预先设定两个阈值 T_1 和 T_2，从而将数据聚为不同的几个簇. Canopy 算法只能大概将数据划分但得到的结果并不够精确，所以更多的是将 Canopy 算法与其他聚类算法结合使用，其算法描述如下：

（Ⅰ）将数据集放入 List 集合，输入阈值 T_1 和 T_2.

（Ⅱ）随机挑选一个点作为中心点，加入到 Canopy 集合中，从 List 集合删除该点.

（Ⅲ）比较 Canopy 集合的点与其他点之间的距离，若小于 T_1，则将这两点聚为相同的一个簇，并从 List 中删除此点；若大于 T_2，则此点加入 Canopy 并从 List 中删除，若大于 T_2 小于 T_1，则依旧在 List 中保存此点.

（Ⅳ）直到 List 集合为空时，算法结束.

2.3 *K*-means 算法

K-means 聚类算法，是一种基于划分的聚类算法，通过输入 k 值从而将数据划分为 k 个簇. *K*-means 算法具有简单、快速等优点，但是算法受到 k 值的影响较大，不同的初始点的选择往往使得聚类的结果产生较大的差异，其算法描述如下：

（Ⅰ）从数据集合中随机选取 k 个值作为初始中心点.

（Ⅱ）计算其他点与初始中心点的距离，归入到距离最近的中心点所对应的簇中.

（Ⅲ）更新簇的中心点，计算并选取新的中心点.

（Ⅳ）重复迭代，直到满足阈值时退出.

K-means 算法的阈值由下式得出

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - \overline{x_i}|^2 \tag{2}$$

式中，E 是所有点的平方误差的总和，x 是集合中每一个点，$\overline{x_i}$ 是簇 C_i 的平均值.

2.4 余弦距离

为了比较两个向量之间的相似度，我们需要求得它们之间的距离，针对文本向量的特点，用余弦距离作为两个向量的相似度，两个 n 维向量 $d(d_1, d_2, \cdots, d_n)$ 和 $c(c_1, c_2, \cdots, c_n)$ 之间的余弦距离表示为 sim，则 sim 的值为

$$\mathrm{sim} = 1 - \frac{(d_1c_1 + d_2c_2 + \cdots + d_nc_n)}{\sqrt{d_1^2 + d_2^2 + \cdots + d_n^2}\sqrt{c_1^2 + c_2^2 + \cdots + c_n^2}} \tag{3}$$

3 PR-HD 并行算法设计

PR-HD 算法主要分为数据预处理与文本向量化、确定聚类中心向量和聚类分析几个部分. 通过 PR-HD 算法，可以把原来庞大且无序的评论数据进行数据规整并从中提取来自产品不同方面的评论热点，从而为产品的生产者、销售者与消费者提供宝贵的意见.

3.1 预处理与文本向量化

PR-HD 算法的预处理与文本向量化阶段就是通过各种方式采集商品评论，剔除无用文本并将评论变成一个由词汇所组成的集合. 为了将文本形式的数据转换成可以通过数字计算的数字形式，对文本进行向量化的操作采用 tfidf 作为文本词汇的权值，这样所有的评论可以转化成＜w_1：tfidf；w_2：tfidf＞的向量格式. 其中，w_i 是单词所对应的词 id，后面紧跟的是单词所占的权值，将所有文章整合为向量格式并传递给下一阶段.

3.2 聚类个数确定

为了找出商品评论数据的大致聚类簇的个数，PR-HD 算法在聚类个数确定阶段首先对上一阶段传递的数据集进行“粗聚类”，使用 Canopy 算法来得出下一聚类所需要的核心聚类中心点，即 k 值.

此阶段的重点是选取合适的阈值 T_1、T_2，我们规定 T_1 大于 T_2 且阈值的选取应当根据实际的情

况进行调整,从而得到更符合预期的结果. 当 T_1 设置偏大时,会造成更多的评论向量属于多个 Canopy,这使得各聚类中心点偏近,各聚类簇差别不大;当 T_1 设置偏小时则造成聚类簇的数量过多,聚类效果较差;当 T_2 设置偏大时,更多的评论向量被标记为强记号,会减少聚类的个数;当 T_2 设置偏小时,会增加簇的个数,算法运行时间也会随之增加. Canopy 阶段的并行化处理机制是各节点将存储在本地上的评论数据集 D_i 生成若干个 Canopy,最后将这些 Canopy 归类汇总,最终得到 k 个聚类.

MapReduce 框架主要由 Map 任务的分配与执行和 Reduce 任务的汇总与执行两部分构成. Canopy 在 Map 阶段,每一个节点随机抽取本台机器中存储的评论向量集 D_i 中的评论向量 v_1 作为一个 Canopy 中心向量,生成一个 Canopy 的集合 canopies,计算其与其他向量距离,使用余弦距离 sim()表示两个向量之间的距离,输出局部中心向量<centerid,vector>,Map 阶段每个节点执行的 Map 任务描述如下:

Input: List<vector>D_i,T_1,T_2
Output: < 本地中心向量 id,本地中心向量值>
Begin
① Canopy. add(v_1)
② while D_i ! = null do
③ for each v_i from D_i do
④ if(sim(Canopy. value,v_1)<T_1)
⑤ Canopy. add(v_i)
⑥ if(sim(Canopy. value,v_i)<T_2)
⑦ Delete v_i from D_i
⑧ end for
⑨ for each Canopy from Canopies do
⑩ write(centroid,Canopy. value)
⑪end for
End

Reduce 阶段主要负责将 Map 阶段每个节点输出的本地 Canopy 中心向量进行汇总,并再次执行 Canopy 算法得到全局中心向量,阈值 T_3、T_4 默认等同于 T_1、T_2,输出为<key1,value1>,其 key1 值为最终 Canopy 的 id 值,Value1 为全局中心评论文本向量,Reduce 阶段执行的伪代码描述如下:

Input: List D,T_3,T_4
Output: <全局中心向量 id,全局中心向量值>
Begin
①Canopy. add(v_1)
② while D ! = null do
③for each v_i from D do
④ if(sim(Canopy. value,v_i)<T_3)
⑤ Canopy. add(v_i)
⑥ if(sim(Canopy. value,v_i)<T_4)
⑦ Delete v_i from D
⑧ end for
⑨ for each Canopy from Canopies do
⑩ write(centroid ,Canopy. value)
⑪ end for
End

聚类个数确定阶段解决了无法确定商品评论讨论话题数目的问题,此阶段得到了商品评论数据集的类别数并给出了下一聚类分析阶段的聚类中心.

3.3　聚类分析

PR-HD 算法的聚类分析阶段指首先获取上一阶段输出的评论数据聚类中心,通过 K-means 算法聚类后获取最终的聚类簇,分析各聚类簇中权重较高的词汇从而得出商品评论的热点信息.

K-means 聚类过程是根据得到的聚类中心向量作为 k 值,将遍历后的评论文本向量通过计算距离分别划分到与之距离最近的簇中,这里采用余弦距离 sim 计算评论文本向量之间的相似度.

K-means 算法的 Map 阶段,主要任务为逐条读入本地节点中的评论向量集 D_i,计算其与哪个中心点 center[i](初次的中心点集为 List<Canopy>)最接近,便把它聚类到中心向量所对应的簇中,其输出结果的<key1,value1>的 key1 值为簇的 id,value 值为相对应的评论向量. Map 阶段中每一个 Map 任务的描述如下:

Input:List<vector>D,center[i]
Output:< 簇 id,本地评论向量 >
Begin
① for each v_i from D_i do
② double min_sim=∞,dist=0
③ for i=0 to k do
④ dist=sim(v_i,center[i]);
⑤ if(dist<min_sim)
⑥ min_sim=dist
⑦ clusterid=i;
⑧ end for
⑨ write(clusterid,vector)
⑩ end for
End

Reduce 阶段,接收并汇总每一个 Map 任务的输出后,重新计算 id 相同的簇所对应的新的中心评论向量,并作为下一次 Map 阶段的输入. 其输出结果的<key1,value1>键值的 key 值为簇的 id,value

值为新的中心向量. Reduce 阶段的描述如下：

```
Input: clusterid ,List<vector>C
Output:< 簇 id,新的中心评论向量 >
Begin
①double num=values. lenght
②double sim[],ave[]
③for each vi from D do
④for i=0 to vi. length do
⑤  sum[i] + =vector. value[i]
⑥  avg[i] = sum[i]/num
⑦ end for
⑧  write(clusterid,avg)
⑨end for
End
```

将 Reduce 阶段的输出重新作为 Map 阶段的输入，算法进入多次迭代，直到达到预先设定的迭代次数或新的中心向量与原中心向量的距离小于一定的阈值为止.

取出最后一次迭代结果，其内容为簇 id 和每一个簇中所包含的评论文本向量数据，再取出每一个簇中所有向量维度上权值最高的词，得到这个簇中的重点信息，通过分析所有簇的重点词汇，可以得出关于这个商品不同方面的评论重点，从而实现了获取商品评价热点的目的.

4 实验结果与分析

4.1 实验环境

实验使用了 3 个节点所构成的 Hadoop 集群，每个结点所使用的计算机配置均完全相同，其配置如下：CPU 为 Intel(R) Core(TM)i7-7700，核心数为 4，主频 3.6 GHz；内存大小为 16 GB；每台计算机均安装了 Ubuntu 16.04 系统；Hadoop 版本为 2.2.0；JDK 使用了 1.8.0 版本.

4.2 评价标准

(Ⅰ) 准确率

一个算法是否能有效解决问题往往需要对其进行正确性的检测，为了检测 PR-HD 算法成功将获取到的评论数据按不同的方面提取热点的可行性，我们引入准确率作为本算法的评论指标.

通过判断 PR-HD 算法是否把应当属于同一类的两个评价向量聚到相同的簇类，来考量它的准确性. 首先需要根据各评论文本向量对之间的关系构建如表 1 所示的混淆矩阵.

表 1 混淆矩阵

Tab. 1 Confusion matrix

类别	同类	不同类
同簇	TP	FP
不同簇	FN	TN

TP 为同类且聚为相同簇的个数；FP 为不同类但聚为相同簇的个数；FN 为相同类却聚为不同簇的个数；TN 为不同类且不同簇的个数.

根据以上混淆矩阵的 4 个值，可以求得算法的准确率. 设评论向量共有 n 个，则共产生评论向量对 C_n^2 个，混淆矩阵中各元素的关系如下：

$$c_n^2 = \mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN} \tag{4}$$

算法的积极正确率 PA 表示为相同类的评论向量对中被正确聚类的评论向量对所占的比值，可以由下式得到

$$\mathrm{PA} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

算法的消极正确率 NA 表示不同类的评论向量对中被正确聚类的评论向量对所占的比值，可以由下式得到

$$\mathrm{NA} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{6}$$

综合积极正确率与消极正确率的值后，我们使用平均正确率 AA 作为算法的最终评价，AA 的值可以由下式得到

$$\mathrm{AA} = \frac{\mathrm{PA} + \mathrm{NA}}{2} \tag{7}$$

平均正确率 AA 代表着整体评论向量聚类后的正确性，平均正确率越高，代表着聚类越精确. 正常情况下，我们希望可以尽可能地提高算法的平均准确率 AA 的值.

(Ⅱ) 加速比

加速比通常被用来衡量一个算法并行化的效果，设一台计算机完成一个串行算法所需要的时间为 T_s，多台计算机并行完成一个算法所需要的时间为 T_p，则加速比 S 为

$$S = \frac{T_S}{T_P} \tag{8}$$

4.3 结果分析

4.3.1 准确率分析

(Ⅰ) PR-HD 算法的准确率分析

本文使用的数据集是通过爬虫爬取的某互联网

电商平台手机类评论数据集共 96 万条评论. 为保证测试算法的准确率,我们从中随机抽取来自屏幕、外观、续航、拍照、系统和物流 6 个方面为评论主题的评论共 60 条来检验算法的正确性. 对归属于 6 个评论热点的 60 条评论数据进行人为标注. 由于采用余弦距离计算文本相似度,所以阈值的选取为 0-1 之间. 根据实验数据集共 6 个方面这一实际情况对阈值进行调整,我们最终将 T_1 设置为 0.9,T_2 设置为 0.7,得到了更符合我们找出 6 个评论热点的预期结果. 最终,PR-HD 算法通过给定阈值自主计算出评论热点话题为 6 个,并给出 6 个聚类中心点,依据 6 个聚类中心点进行聚类的结果构建出的混淆矩阵如表 2 所示.

表 2　PR-HD 算法混淆矩阵

Tab. 2　PR-HD algorithm confusion matrix

类别	同类	不同类
同簇	282	20
不同簇	18	1480

由表 2 可以看出,60 条确切分组的评论数据集一共可以产生 1 770 个评论向量对数. 其中,TP 值为 282,FP 值为 20,FN 值为 18,TN 值为 1 480,将上述数值分别代入到公式(5)～(7)中,求得的算法结果的准确率如表 3 所示:

表 3　PR-HD 算法准确率

Tab. 3　PR-HD algorithm accuracy

准确率	PA	NA	AA
	94%	98.6%	96.3%

分析表 3 可知,算法的积极准确率为 94%,相对应的算法消极准确率为 98.6%,结合两者,最终得到的算法平均准确率为 96.3%. 由此可以得出,PR-HD 算法可以准确地找出商品评论中的重点信息.

(Ⅱ) 相关工作对比

模糊聚类(fuzzy c-means,FCM)算法是一种基于划分的聚类算法,同样可以被应用于商品评论的热点挖掘. 将 PR-HD 算法与基于 FCM 算法的商品评论热点发现算法进行对比,使用测试数据对比两种算法的正确率. 通过输入明确的评论热点个数 6 后,基于 FCM 的商品评论热点发现算法的聚类结果构建出如表 4 所示的混淆矩阵.

表 4　FCM 算法混淆矩阵

Tab. 4　FCM algorithm confusion matrix

类别	同类	不同类
同簇	275	30
不同簇	25	1 470

同理,使用公式(5)～(7)可以分别求出基于 FCM 的商品评论热点发现算法的积极准确率、消极准确率和平均准确率、结果如表 5 所示.

表 5　FCM 算法准确率

Tab. 5　FCM algorithm accuracy

准确率	PA	NA	AA
	94%	98.6%	96.3%

由表 5 可知,算法的积极准确率为 91.6%,算法消极准确率为 98%,最终得到的算法平均准确率为 94.8%,将 PR-CH 算法与基于 FCM 的商品评论热点发现算法的准确率进行对比可以得到表 6.

表 6　准确率对比

Tab. 6　Comparison of accuracy

算法	准确率		
	PA	NA	AA
PR-HD	94%	98.6%	96.3%
FCM	91.6%	98%	94.8%

由表 6 可知,在提取商品评论热点方面,PR-HD 算法的准确率高于基于 FCM 的商品评论热点发现算法的准确率,这也说明 PR-HD 算法在商品评论的热点发现方面具有更好的效果.

4.3.2　加速比分析

为了衡量 PR-HD 算法的并行效果,本文选用 3 种不同规模的评论数据集并提前完成了去杂、分词等操作,得到的格式化数据规模分别为 321.7 MB、553.3 MB 和 1.34 GB. 分别测试 3 种规模数据在 1、2 和 3 个集群节点上的运行时间,所得结果如表 7 所示.

表 7　算法运行时间(s)

Tab. 7　Algorithm runtime(s)

数据量/MB	1 个节点	2 个节点	3 个节点
321.7	902.75	544.01	423.063
553.3	1 468.66	834.12	616.06
1 340	2 879.16	1 580.94	1 127.011

根据表 7，不同规模的数据在不同节点的加速比可以用公式(8)求出，如表 8 所示.

表 8 算法运行加速比

Tab. 8 Algorithm running speedup

数据量/MB	2 个节点	3 个节点
321.7	1.65	2.13
553.3	1.76	2.38
1 340	1.82	2.56

由表 8 可知，在数据规模保持不变的情况下，当我们增大集群的节点数目时，整个集群的总体性能也会随之增强. 当数据节点不变时，随着评论数据集规模的增大，加速比也会随之增高. 加速比曲线如图 1 所示.

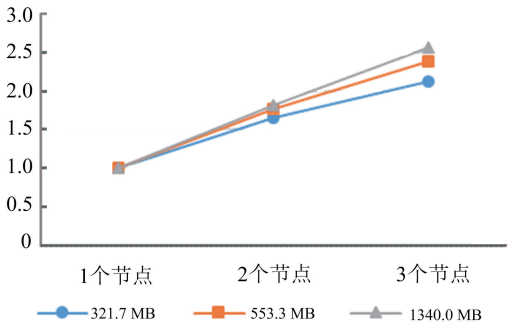

图 1 不同规模数据量的加速比

Fig. 1 Accelerating ratios of data at different scales

从图 1 可以看出，随着数据量的增大，加速比的曲线更加接近线性. 实验结果表明，PR-HD 算法在处理较大规模的数据时展现出了良好的性能，可以有效地提高算法的执行效率，并且具有比较好的加速比.

在使用 1.34 GB 规模评论数据的前提下，将 PR-CH 算法与基于 FCM 的商品评论热点发现算法进行加速比的对比可以得到表 9.

表 9 加速比对比

Tab. 9 Acceleration ratio comparison

算法	2 个节点	3 个节点
PR-HD	1.82	2.56
FCM	1.80	2.53

分析表 9 可以得出，虽然基于 FCM 算法的商品评论热点发现算法同样具有不错的加速比，但 PR-HD 拥有着更高的准确率，因此 PR-HD 算法在超大规模评论数据的热点发现上拥有着更好的效果.

5 结论

本文提出了一种基于 MapReduce 的商品评论热点发现算法——PR-HD 算法，在多节点的集群环境下，对算法进行了实验. 实验证明，算法具有较高的准确性，同时在集群性能增加时明显提高算法的执行效率，可以处理超大规模的数据量. 对算法所得到的结果进行分析，能够较好地实现商品评论热点提取，得到商品在不同方面的评论反馈信息，实现了商品评价的深度价值挖掘，对商品开发人员到消费者均具有很高的参考价值.

另外，本文在算法执行过程中，虽然避免了人为规定簇的个数所带来的缺陷，提升了商品评论挖掘的质量，但是却引入了新的阈值的概念，在某些情况下，阈值大小的设定又会成为新的影响结果的因素. 下一步的研究是如何通过设计新的算法来根据评论自动选择合适的阈值，进一步减少人为因素的干扰，提高热点发现的准确率.

参考文献(References)

[1] 刘建红. 基于 Hadoop 平台的聚类算法并行化研究[D]. 长春：吉林大学，2017.
[2] 郑淑蓉，吕庆华. 中国电子商务 20 年演进[J]. 商业经济与管理，2013，(11)：5-16.
[3] 伍星，何中市，黄永文. 产品评论挖掘研究综述[J]. 计算机工程与应用，2008，44(36)：37-41.
[4] 罗慧钦，陆向艳，张雄宝，等. 基于隐朴素贝叶斯的商品评论情感分类方法[J]. 计算机工程与设计，2017，38(1)：203-208.
[5] 董祥和. 基于情感特征向量空间模型的中文商品评论倾向分类算法[J]. 计算机应用与软件，2016，33(8)：319-322.
[6] 王刚，杨善林. 基于 RS—SVM 的网络商品评论情感分析研究[J]. 计算机科学，2013，40：274-277.
[7] 王丹丹，祖颖，朱平. AABC-SVM 模型及其在商品评论情感分类中的应用[J]. 计算机应用与软件，2017，34(9)：33-37.
[8] 李斌. 基于聚类挖掘技术在电子商务网站中的应用[J]. 电脑知识与技术，2014(5)：1147-1149.
[9] 罗好. 聚类数据挖掘在商场中的应用及 K-means 聚类算法改进研究[D]. 重庆：重庆大学，2005.
[10] 郑丹，王潜平. K-means 初始聚类中心的选择算法[J]. 计算机应用，2012，32(8)：2186-2188.
[11] 邱荣太. 基于 Canopy 的 K-means 多核算法[J]. 微计算机信息，2012，28(9)：486-487.
[12] 唐晨馨. 基于层次化 AP 聚类的商品评论数据标签化[D]. 江门：五邑大学，2017.