文章编号：2096-1472(2018)-05-01-03
DOI:10.19644/j.cnki.issn2096-1472.2018.05.001

大数据架构下的热词发现与可视化技术研究

胡瑞娟
(信息工程大学，河南 郑州 450000)

摘　要：在大数据背景下，数据膨胀的速度已经远远超出了人工分析的能力范围，因此，如何在大数据时代构建热词发现与可视化机制尤为紧迫和重要。本文通过研究Hadoop大数据平台下的MapReduce计算框架和TF−IDF算法，给出了TF−IDF算法在Hadoop分布式并行化计算平台下的具体实现，并以此并行化算法作为大数据架构下热词发现技术的核心算法，然后利用可视化工具对结果进行分析处理。结果表明，TF−IDF并行化算法可以较好地发现大规模数据量中的热点词汇；与传统单机下的算法相比，该算法处理效率更高。

关键词：Hadoop；TF−IDF并行化；热词发现；可视化
中图分类号：TP391　**文献标识码**：A

Research on Hot Word Discovery and Visualization Technology Based on Big Data Architecture

HU Ruijuan
(*Information Engineering University*,*Zhengzhou* 450000,*China*)

Abstract: The speed of data expansion is far beyond the ability of artificial analysis in the era of big data.Therefore,it is particularly urgent and important how to build hot word discovery and visualization mechanism.By studying the MapReduce computing framework and TF-IDF algorithm under the Hadoop platform,this paper gives the concrete implementation of the TF-IDF algorithm under the Hadoop distributed parallel computing platform,and uses this parallel algorithm as the core algorithm of the hot word discovery technology based on the big data architecture,and then uses the visualization tool to display and analyze the results.The results show that the TF-IDF parallelization algorithm can find the hot words in large amount of data much better.Compared with traditional single-machine algorithms,this algorithm is more efficient.

Keywords: Hadoop;TF-IDF parallelization;hot word discovery;visualization

1 引言(Introduction)

大数据时代，数据膨胀的速度已经远远超出了人工分析的能力范围。要从这些庞杂的数据中获取有效信息更是难上加难。如何能够简单、快速并且有效地得知人们感兴趣的内容是一个重要的问题。热词发现作为近年来的一个研究热点，不失为一个切实可行的解决方案。因为热词[1]往往能够反映一个国家、一个地区在一个时期人们普遍关注的问题和事物，具有时代特征，所以及时准确地发现当前热点信息并进行整理分析对于了解民意动向、分析舆情走势十分重要。其结果不仅为政府和有关机构及时了解并处理重大社会热点问题提供很大的参考价值，也为广大人民群众能够在信息爆炸时代里了解社会热点提供了快速而有效的方法[2]。

本文通过研究Hadoop大数据平台下的MapReduce计算框架和TF−IDF算法，给出了TF−IDF算法在Hadoop分布式并行化计算平台下的具体实现，并以此并行化算法作为大数据架构下热词发现技术的核心算法，然后利用可视化工具R语言[3]对结果进行分析处理。

2 大数据与Hadoop平台(BigData and Hadoop)

IDC公司从四个特征定义大数据，即海量的数据规模(Volume)、快速的数据流转和动态的数据体系(Velocity)、多样的数据模态(Variety)和巨大的数据价值(Value)。根据大数据的生命周期，大数据的技术体系可以分为大数据采集与预处理[2]、大数据存储与管理、大数据计算模式与系统、大数据分析与挖掘、大数据可视化计算和大数据隐私安全等方面。我们通常选用Hadoop系统来存储、管理、分析这些数据，以获取更多有价值的信息[4]。

Hadoop是Apache软件基金会旗下的一个开源分布式计算平台，以Hadoop分布式文件系统HDFS和MapReduce为核心的Hadoop为用户提供了系统底层细节透明的分布式基础架构。HDFS的高容错性、高伸缩性允许用户将Hadoop部署在低廉的硬件上，形成分布式系统；由于Hadoop拥有可计量、成本低、高效、可信等特点，基于Hadoop的应用已经开始遍布互联网领域。MapReduce分布式编程模型允许用户在不了解分布式系统底层细节的情况下运行应用程序[5]。

3 大数据架构下热词发现理论基础(Theoretical foundation of hot word discovery based on bigdata architecture)

3.1 MapReduce计算框架

分布式文件系统(HDFS)和MapReduce编程模型是Hadoop的主要组成部分。MapReduce模型[6]的计算流程如图

基金项目：2017年度洛阳市社会科学规划项目，大数据架构下的热词发现与可视化技术研究.

2所示。分布式文件系统主要负责各节点上的数据的存储，并实现高吞吐的数据读写。MapReduce计算模型的核心部分是Map和Reduce两个函数[7]。Map的输入是in_key和in_value，指明了Map需要处理的原始数据。Map的输出结果是一组<key,value>对。系统对Map操作的结果进行归类处理。Reduce的输入是(key，[value1…value m])。Reduce的工作是将相同key的value值进行归并处理最终形成(key，final_value)的结果，所有的Reduce结果并在一起就是最终结果。

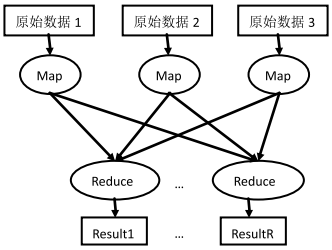

图1 MapReduce模型的计算流程

Fig.1 Calculation process of MapReduce

3.2 TF-IDF热词发现算法

TF-IDF(Term Frequency-Inverse Document Frequency)[8]是一种服务数据挖掘的常用的加权技术。该算法可以评价一个单词对于语料库中的某一文件的重要程度。该算法的计算公式为：

$$TF-IDF = TF \times IDF \tag{1}$$

其中，TF(词频，Term Frequency)指的是某一个给定的词语在该文件中出现的频率。同一个词语在长文件里可能会比短文件有更高的词数，而不管该词语重要与否。对于在某一特定文件里的词语t_i来说，它的重要性可表示为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

其中，$n_{i,j}$是该词t_i在文件d_j中的出现次数，而分母则是在文件d_j中所有字词的出现次数之和。

IDF(逆向文件频率，Inverse Document Frequency)是一个词语普遍重要性的度量。某一特定词语的IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到：

$$idf_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|} \tag{3}$$

其中，$|D|$：语料库中的文件总数，$|\{j:t_i \in d_j\}|$：包含词语t_i的文件数目(即$n_{i,j} \neq 0$的文件数目)如果该词语不在语料库中，就会导致被除数为零，因此一般情况下使用$1+|\{j:t_i \in d_j\}|$。

可以看到，TF-IDF与一个词在文档中出现的次数成正比，同时与该词在整个语料库中出现的次数成反比。因此，要想得到TD-IDF值，需要分别求出TF、IDF两个值。无论是计算TF还是IDF的值，由于文档总数是一个常量，所以在得到TF后，只需要统计所有语料中包含该单词的文档的个数，即可进行TF-IDF值的计算。从TF-IDF的计算公式中不难看出，它非常适合利用分布式并行化平台进行计算求解。

3.3 R语言

R语言是一个拥有强大的统计计算和制图能力的可视化工具。作为一个统计分析软件，R集统计分析和图形显示于一体，与其他统计分析软件相比，R还具有开源且免费使用、与各种OS兼容性好、丰富的插件支持、互动性好等特点[9,10]。

4 大数据架构下的热词发现与可视化系统的构建(Construction of hot word discovery and visualization system based on bigdata architecture)

根据MapReduce、TF-IDF算法等基本理论，本文设计了大数据架构下的热词发现系统，在Hadoop平台上实现了TF-IDF并行化算法。该系统主要包括数据采集(获取实验语料)、数据预处理、Hadoop平台搭建、TF-IDF算法并行化实现及结果可视化五部分。系统总体框架如图2所示。

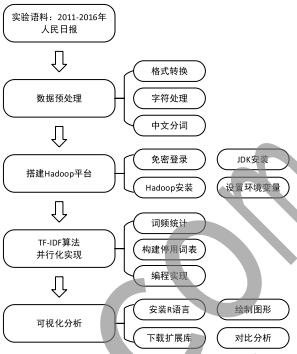

图2 热词发现系统框架

Fig.2 Hot word discovery system framework

4.1 数据采集与预处理

(1)实验数据

本文的实验数据是从人民日报官网http://paper.people.com.cn上，利用网络爬虫获取2011年至2016年这六年间每一天的人民日报电子版，得到格式为PDF，总大小为34.3GB，共计2191个文件。本设计以上述文件作为实验的原始数据。

(2)数据预处理

首先对爬取到的人民日报进行格式转换，格式化文件名(文件名中不得含有中文及中文字符)，修改文档的编码方式，并去除所有的特殊符号，为接下来的中文分词做准备。然后利用斯坦福大学NLP研究室提供的Stanford CoreNLP工具包对语料进行中文分词。最终将数据集按年度分开(years)，每年分为12个文档(每月一个文档，month)。

4.2 TF-IDF算法的并行化实现

利用Hadoop集群环境及MapReduce计算框架，对TF-IDF算法进行并行化实现，其核心思想是将任务进行分割，并分配到集群主机运行。通过分割数据，可以将统计文档中单词词频的任务并行化处理，从而减少处理时间，提高处理效率。

具体流程如下：

(1)计算各文档中单词的词频TF

在读入目标文档所在的HDFS路径后，将原始数据进行分片并传给Map函数，在Map中利用正则表达式删除空行和空字符串，然后传给Reduce函数统计单词词频，并将结果输出到临时文件中保存，作为下一个MapReduce的输入。该函数设计如下：

Map()：

Input:<filepath>

Output:<filename wordname,1>

Reduce()：

Input:<filename wordname,1>

Output:<filename wordname,tf>

(2)计算单词的逆文档频率IDF

读取上一步输出的临时文件tfidf-tf作为本次Map函数的

输入。在Map中拆分单词，按频次为1，以便Reduce分文件名进行合并。在Reduce中统计所有语料中包含wordname一词的文档总数termdocnum和文档总数alldoc，然后按照公式$idf=\log[alldoc/(termdocnum+1)]$计算idf值，并将结果以键值对<wordname,idf>的形式输出到临时文件中保存，作为下一个MapReduce的输入。该函数设计如下：

Map()：

Input:<filename wordname,tf>

Output:<filename wordname,filenum>

Reduce()：

Input:<filename wordname,filenum>

Output:<wordname,idf>

(3)计算单词的TF-IDF

利用两个Map函数分别读取前两步输出的临时文件tfidf-tf和tfidf-idf作为输入。在Reduce中按照公式$tfidf=tf*idf$计算tfidf值，并将结果以键值对<filename wordname,tfidf>的形式输出到结果文件tfidf-tfidf中保存，作为结果分析的依据。该函数设计如下：

Map()：

Input:<filename wordname,tf>

<wordname,idf>

Output:<wordname,tf>

<wordname,idf>

Reduce()：

Input:<wordname,tf>

<wordname,idf>

Output:<filename wordname,tfidf>

计算TF-IDF的整个流程如图3所示。

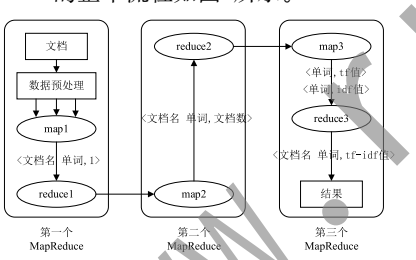

图3 TF-IDF计算流程

Fig.3 Calculation process of TF-IDF parallelization

与单机环境下运行TF-IDF算法进行对比，效率提高了很多。

4.3 利用R语言进行结果可视化

以数据集years作为实验语料，在集群模式下运行程序，得到结果后利用R语言分别对2011—2016年的年度热词进行可视化展示，2016年度热词可视化结果如图4所示。

图4 2016年年度热词词云

Fig.4 2016 annual hot word cloud

可以看出，2016年度的热点词汇有“两学一做”“三严三实”“供给侧结构性改革”“大数据”“党风廉政建设”“创新”“改革”“全面二孩”“战略支援部队”“深化国防军队改革”“工匠精神”“强军目标”“军事斗争准备”“人工智能”“营改增”等。而互联网中给出的2016年年度热词为：“供给侧改革”“全面二孩”“工匠精神”“人工智能”“营改增”“不忘初心”“里约奥运会”“电信诈骗”“房价”“网约车”。考虑到本实验的语料来源仅为《人民日报》，其中并不完全包含互联网中的内容，而互联网上公布的结果则是综合多方面内容得到的，因此二者的结果存在一定的出入。但从整体上来看，由该问题导致的误差在可接受范围内，实验结果较为可靠，该实验能够较好地完成大数据规模下的热词发现任务。

5 结论(Conclusion)

大数据架构下的热词发现与可视化技术研究，可以帮助相关人员及时准确地发现当前热点信息，这些信息为政府和有关机构及时了解并处理重大社会热点问题提供很大的参考价值；另一方面，通过对热词发现结果的可视化，直观形象地展示出热词，可以为广大人民群众能够在信息爆炸时代里了解社会热点提供了快速而有效的方法；同时，一些优秀的网络热词可以作为汉语言文库的一种有益补充和丰富，为语言学家研究汉语言提供一定的帮助。

参考文献(References)

[1] 耿升华.新词识别和热词排名方法研究[D].重庆大学,2013.

[2] 刘晨,焦合军.基于HADOOP集群的数据采集和清洗[J].软件工程,2016,19(11):20-24.

[3] 汤银才.R语言与统计分析[M].高等教育出版社,2008.

[4] 崔文斌,牟少敏,王云诚,等.Hadoop大数据平台的搭建与测试[J].山东农业大学学报自然科学版,2013,44(4):550-555.

[5] 孟永伟,黄建强,曹腾飞,等.Hadoop集群部署实验的设计与实现[J].实验技术与管理,2015,32(1):145-149.

[6] 郝树魁.Hadoop HDFS和MapReduce架构浅析[J].邮电设计技术,2012(7):37-42.

[7] Luis Torgo,李洪成,陈道轮,等.数据挖掘与R语言[J].计算机教育,2013(12):102-102.

[8] 张建娥.基于TFIDF和词语关联度的中文关键词提取方法[J].情报科学,2012(10):110-112;123.

[9] 吴丹露,魏彤,许家清.R语言环境下的文本可视化及主题分析——以社会服务平台数据为例[J].宁波工程学院学报,2015,27(1):19-25.

[10] 叶文春.浅谈R语言在统计学中的应用[J].中共贵州省委党校学报,2008(4):123-125.

[11] 使用Stanford CoreNLP工具包处理中文[EB/OL].http://blog.csdn.net/u014749291/article/details/51152007,2016-04-14.

[12] 使用Stanford NLP进行中文分词[EB/OL].http://blog.sectong.com/blog/corenlp_segment.html,2016-03-26.

[13] 赵伟燕.基于Hadoop平台的TFIDF算法并行化研究[D].内蒙古科技大学,2013.

作者简介：

胡瑞娟(1983-)，女，硕士，讲师.研究领域：大数据智能处理.