

# 基于 Hadoop 的网络舆情监控平台设计与实现

李 晨,杨子江,朱世伟,于俊凤

(山东省科学院 情报研究所,山东 济南 250014)

**摘 要:**文中设计并实现了一种基于 Hadoop 的网络舆情监控系统。该系统以 HDFS 作为底层存储系统,在其上构建基于 HBase 的分布式数据库对舆情信息进行统一存储管理。首先利用基于 MapReduce 的分布式网络爬虫进行数据抓取,以解决单机爬虫效率低、可扩展性差等问题;其次采用 Canopy 结合  $K$ -means 的二次聚类算法,克服单一  $K$ -means 聚类算法的不足,以提高文本聚类的效率和准确度;最后实现基于查询的话题追踪策略,对热点话题进行有效跟踪分析。仿真实验表明:Canopy-Kmeans 聚类方法比传统  $K$ -means 方法漏报率、误报率分别降低 1.24%、0.09%,最小标准代价降低 1.681%。系统通过提供可视化舆情分析报告,为企业或单位及时掌握舆情热点、制定舆情策略提供科学、系统的技术支持。

**关键词:**Hadoop;MapReduce;舆情监控;文本聚类;热点发现;话题跟踪

中图分类号:TP311.1

文献标识码:A

文章编号:1673-629X(2016)02-0144-06

doi:10.3969/j.issn.1673-629X.2016.02.033

## Design and Implementation of Network Consensus Monitoring System Based on Hadoop

LI Chen, YANG Zi-jiang, ZHU Shi-wei, YU Jun-feng

(Information Institute, Shandong Academy of Sciences, Jinan 250014, China)

**Abstract:** A network consensus monitoring system based on Hadoop was designed and realized. The system adopts HDFS as the underlying storage system, and then it builds a distributed database based on HBase with it to realize unified storage and management on the network consensus information. Firstly, it grabs the data with the distributed web crawler based on MapReduce to solve the problems of low efficiency and poor expansibility of single crawler. Then it uses the secondary clustering algorithm with Canopy combined with  $K$ -means, which can overcome the shortages of single  $K$ -means clustering algorithm and could improve the efficiency and precision of text clustering. Finally, it could realize the topics tracking strategy based on query, also could be effective track and analysis of hot topics. The simulation experiment results show that compared with the traditional methods, the false negative and false positive of Canopy-Kmeans clustering method is lower at 1.24% and 0.09% respectively, the minimum standard price is lower at 1.681%. Through providing the visualized analysis of network consensus, the system proposed could provide scientific and systematical technology support for enterprises and scientific institutions to learn the hot network consensus and make network consensus strategy.

**Key words:** Hadoop; MapReduce; monitoring public opinion; text clustering; hot topic founding; topic tracking

### 1 概 述

随着信息技术以及互联网的快速发展,其产生的海量、异构、动态的新闻数据使得人们很难快速、高效地找到用户感兴趣的新闻。如何对这些新闻数据进行准确地挖掘与分析,实现对新闻话题的持续追踪和舆情预测已成为目前舆情分析中一个极其重要的研究方向。

传统的话题追踪和舆情监控系统通常是基于昂贵的工作站或者服务器集群<sup>[1]</sup>,采用流量镜像的方法监

控信息源,并结合传统的数据挖掘算法对获取的数据进行文本统计与分析。基于流量镜像方法虽然可以比较全面地收集各种网络信息,但也存在成本高、可扩展性差以及容易产生单点通信故障等问题。同时,由于互联网信息的爆炸式增长,产生海量的网络信息,如何存储并处理这些海量、异构的非结构化信息便成了一个新的研究课题。Hadoop<sup>[2]</sup>的产生为这一课题提供了有效的解决方案。Hadoop 技术对海量信息的存储与处理提供了高效、可靠、可扩展的解决办法。

收稿日期:2015-04-24

修回日期:2015-08-03

网络出版时间:2017-01-04

基金项目:山东省科学院青年基金项目(2013QN036);山东省科技发展计划(2013GGX10127,2014GGX101013)

作者简介:李 晨(1988-),男,实习研究员,研究方向为数据挖掘、大数据分析。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20160104.1505.038.html>

在此基础上,文中设计了一个基于Hadoop的网络舆情监控系统。该系统采用分布并行方式对互联网信息进行采集,采用基于查询的技术进行话题追踪和舆情监控,通过基于Mahout<sup>[3]</sup>实现的朴素贝叶斯算法对新闻话题进行分类;并使用基于Canopy算法<sup>[4]</sup>和K-means算法<sup>[5]</sup>相结合的聚类方法对新闻话题进行层次聚类,最后结合报道数量、来源以及报道速度等因素实现对新闻热点话题的量化与跟踪。通过将该系统应用到山东省科学院舆情分析平台,验证了该系统可以有效地实现对新闻话题的追踪和网络舆情的预测。

## 2 研究现状和意义

网络舆情监控平台主要是针对海量数据进行网络舆情分析<sup>[6]</sup>,其中涉及到的关键内容有:网页信息获取、文本分类、文本聚类、热点和敏感话题发现以及话题跟踪等内容。对此,国内外的主要研究现状如下:

文本分类方面,SVM的产生是近年来文本分类领域最重要的进展之一,虽然SVM在大规模数据集上的训练收敛速度较慢,但是它的分隔面模式有效地克服了样本分布、冗余特征以及过拟合等因素的影响,具有很出色的泛化能力,有文献已经指出SVM在效果和稳定性上具有相当的优势。Pang(2002)等学者比较支持向量机(SVM)、最大熵(ME)、朴素贝叶斯(NB)等算法在语义倾向上的文本分类效果,实验结果发现,SVM的文本分类准确程度达到约80%,属于较好的一种。

聚类方面,为了适应语料增长的特性,卡内基梅隆大学和BBN公司分别使用Single-pass算法和增量K-means算法进行聚类。宾夕法尼亚大学在计算文档之间的相似度时使用IDF加权的余弦值,以此来提高聚类质量<sup>[7]</sup>。Dragon为解决在线识别问题,提出一种利用K-means聚类方法中的第一个迭代过程来确定报道所属话题类<sup>[8]</sup>。IBM采用两层聚类策略进行话题分析。热点发现采用TDT(话题检测与追踪)中话题检测的相关技术,它主要以原始新闻语料作为研究对象,计算时通过将相关参数进行量化得到最终结果,量化的参数一般有话题的报道频率、话题的分布率、话题的时间属性等<sup>[9]</sup>,还有一些将报道点击、评论、来源等作为计算参数进行量化。

话题跟踪方面,可以分为两种跟踪策略:基于知识的话题跟踪和基于统计的话题跟踪。基于知识的话题跟踪策略比较有名的是Watanabe基于日语新闻广播所研发的话题跟踪系统。基于统计的话题跟踪策略,比较有代表性的是基于分类的话题跟踪,如卡内基梅隆大学在对话题跟踪评测中使用决策树和KNN算法。

在舆情分析研究过程中国内外也产生了很多优秀的软件产品,比如:国外有Dave等研发的ReviewSeer, Liu等研发的Opinion Observer, Gamon等研发的Pulse系统, Wilson等研发的Opinion Finder。国内有北大方正的智思舆情预警辅助决策支持系统,中科院实施的天网工程, TRS公司研发的网络舆情情报监控体系,等等。

文中系统在现有工作研究的基础上,针对新闻网页、论坛、博客以及微博进行数据挖掘研究,利用Hadoop技术平台,将网络舆情信息采用HDFS<sup>[10]</sup>进行存储,并通过MapReduce<sup>[11]</sup>编程模式进行分析,实现快速发现舆情热点,并进行情感分析与热点话题跟踪,为后期的进一步监控提供基础信息保障。

## 3 基于Hadoop的网络舆情监控系统实现

Hadoop包含大量工具,这些工具可以协同工作,来完成多种任务。Hadoop可以归类成一个完整的生态系统,包含大量的组件,从数据存储到集成、数据处理及数据分析等。HDFS作为Hadoop生态系统的基础组件,它可以将大量数据分布到计算机集群之上,实现一次写入,多次读取。Hadoop的主要执行框架是MapReduce,它是一个用于分布式并行数据处理的编程模型。HBase<sup>[12]</sup>是一个构建于HDFS之上的面向列的NoSQL数据库,提供对海量数据的快速读写能力,它利用Zookeeper作为自己的分布式协调工具。Oozie作为一个可扩展的工作流系统,可以协调多个作业的执行。更高层的抽象Pig和Hive可以完成数据分析和类似SQL的数据查询。

文中设计的基于Hadoop的网络舆情监控系统主要包括四个模块:分布式信息采集、信息存储、云分析以及舆情信息展示。分布式信息采集作为系统核心组件主要工作是抓取互联网信息,对抓取的互联网信息进行处理,然后进行存储,为上层分析提供数据支持。系统采用HDFS作为底层数据存储介质,在其之上构建更高层次的HBase和Hive进行数据管理。云分析采用分布式编程设计对原始网页信息进行处理,包括:文本分类、文本聚类、热点计算以及话题跟踪等。基于云计算的网络舆情监控系统总体架构见图1。

### 3.1 分布式信息采集

原始信息采集是建设舆情监控系统的首要任务,该系统主要采用分布式网络爬虫进行信息抓取。采用定时定向的信息采集方式,由采集控制器进行统一调度,通过采集控制器读取相应的站点信息,包括:URL、站点模板等信息,之后就可以利用分布式网络爬虫进行信息采集。系统采集架构见图2。

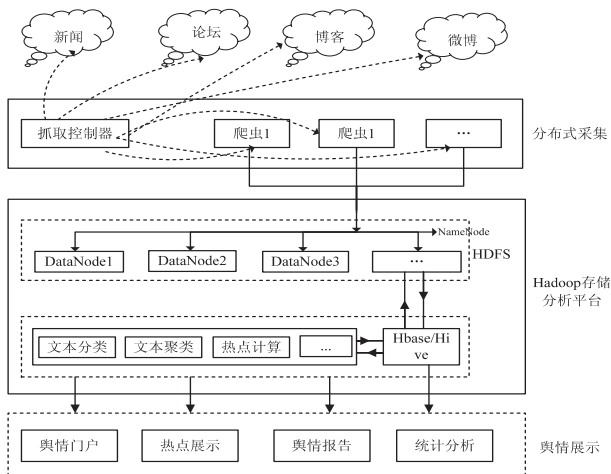


图 1 系统架构图

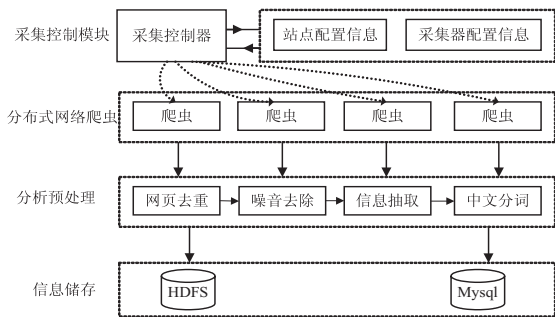


图 2 系统采集架构图

### 3.1.1 网页信息抽取

采集器采集到的原始网页内容往往比较繁杂,可能会包含大量的广告、无用链接以及其他的噪音信息,噪音信息的不便于对网页的内容进行分析,所以首先需要进行网页信息抽取<sup>[13-15]</sup>。在该系统中,采用了基于行块分布函数的通用网页正文抽取方法,该方法无需考虑网页 HTML 结构,无需构建传统 DOM 树,可以从杂乱的网页源码中抽取出有效的、高质量的文本信息。

算法流程为:首先,通过正则表达式去除所有 HTML 标签,只保留文本信息,同时去除标签之后的空白信息也要保留,保留文本成为 Ctext;其次,定义行块。以 Ctext 中的行号为轴,取其周围  $K$  行,称为一个行块 Cblock,行块  $i$  是以 Ctext 中行号  $i$  为轴的行块;再次,定义行块长度。求出一个 Cblock 行块中去除所有空白符后的字符总数作为该行块的长度;然后,定义行块分布函数。以 Ctext 每行为轴,共有  $\text{LinesNum}(\text{Ctext}) - K$  个 Cblock,做出以  $[1, \text{LinesNum}(\text{Ctext}) - K]$  为横轴,以其各自的行块长度为纵轴的分布函数;最后,分析行块分布函数分布区域从而获得网页正文信息。

### 3.1.2 网页信息预处理

网页信息预处理模块包括:分词分词、停用词过滤、文本特征选择等。系统采用具有词性标注功能的 FudanNLP 进行文本分词,为了降低维度,在分词的过

程中只选取名词、动词以及形容词作为最终的分词结果。处理完之后,将原始网页信息和分词信息存入原始网页信息库,为云分析提供原始素材。

### 3.1.3 网页抓取流程

系统抓取流程共包括四步:首先获取待抓取 URL 列表,AccessDBDriver 从传统数据库中读取 URL 列表,然后将该列表写入 HDFS 作为 CrawlDriver 的 Map 任务输入。第二步,CrawlDriver 的 Map 任务读取待抓取 URL 列表,将该列表发送到 Reduce 任务,Reduce 任务进行网页信息抓取。第三步,解析网页内容。ParserDriver 进行网页信息抽取和网页信息预处理。第四步,利用 TransferDriver 将解析过的原始网页信息和分词信息存入 Hbase 库中。

## 3.2 云分析

云分析模块包括文本分类、文本聚类、热点话题发现和热门话题跟踪等。文中采用开源 Mahout 系统实现基于 MapReduce 的分布式文本分类和聚类算法。

### 3.2.1 文本分类

文中采用 Mahout 基于朴素贝叶斯<sup>[16]</sup>的文本分类算法,其分类具体流程是:

(1)文本序列化。将原始网页信息转化为 Mahout 可以直接使用的二进制 SequenceFile 文件,该步骤利用 Mahout 的 seqdirectory 命令实现。

(2)序列向量化。将序列化好的 SequenceFile 文件利用 TF-IDF 生成向量空间模型,同时在该操作中指定分词器。该步骤利用 seq2sparse 命令实现。

(3)向量划分。该步骤通过把向量化的文件随机分成两部分,一部分用来训练,另一部分用来测试,使用的命令是 split。

(4)训练并生成模型。用第三步随机分割的训练数据作为 Naive Bayesian 的输入进行训练并生成模型,使用 trainnb 命令完成。

(5)测试训练集。用第三步的测试数据作为输入对生成的模型进行测试,使用 testnb 命令实现。

### 3.2.2 文本聚类

通过聚类可以将一组文章或文本信息进行相似性的比较,将比较相似的文章或文本信息归为同一类簇。文中采用 Canopy 和  $K$ -means 结合的文本聚类方法。

Canopy 算法的基本思想是:将数据集向量化,然后放到一个集合 list 中,同时设定两个距离阈值  $T_1$  和  $T_2$ ,循环从 list 中去取一个点,作为一个聚类中心,放到 centerlist,并从 list 中移除该点,循环从 centerlist 中比较与周围的点与阈值之间的关系,小于最小  $T_1$  阈值,说明两个值相似,放到一个聚类中,并从 list 中移除;如果大于最大阈值  $T_2$ ,那么就单独作为一个聚类中心,并从 list 中移除;否则不加入到各个聚类中心去,



但依然保留在 list 中,迭代直至 list 中元素为 null,算法结束。

$K$ -means 聚类算法是一种基于样本间相似性度量的间接聚类方法。基本思想是初始随机给定  $K$  个簇中心,按照最邻近原则把待分类样本点分到各个簇中。然后按平均法重新计算各个簇的质心,从而确定新的簇心。一直迭代,直到簇心的移动距离小于某个给定的值。传统  $K$ -means 算法在进行聚类时需要事先指定  $K$  值(类别数目),而往往数据集预先不能确定  $K$  值大小,如果  $K$  取的不合理将会使  $K$ -means 算法误差很大。

为了克服这种缺点,文中首先使用 Canopy 算法进行一次聚类,将较小数目的 Cluster 直接去掉有利于抗干扰,然后在每个 Canopy 内使用  $K$ -means 方法进行二次聚类。

### 3.2.3 热点计算

文中综合考虑报道数量、报道来源、报道速度等因素,利用式(1)进行热点话题计算:

$$\text{focus} = r_n * N/R_N + r_d * N / \sum_{i=0}^N (r_d)_i + g \log_{h_n+5c_n+w} (h_n + 5c_n) \quad (1)$$

其中,  $r_n$  为这段时间内相关报道的数量;  $R_N$  为这段时间所有报道的总数;  $N$  为这段时间内话题总数目;  $r_d$  为总报道天数;  $\sum_{i=0}^N (r_d)_i$  为所有话题的报道天数和;  $h_n$  为话题点击数;  $c_n$  为话题评论数。

具体实现流程见图3。

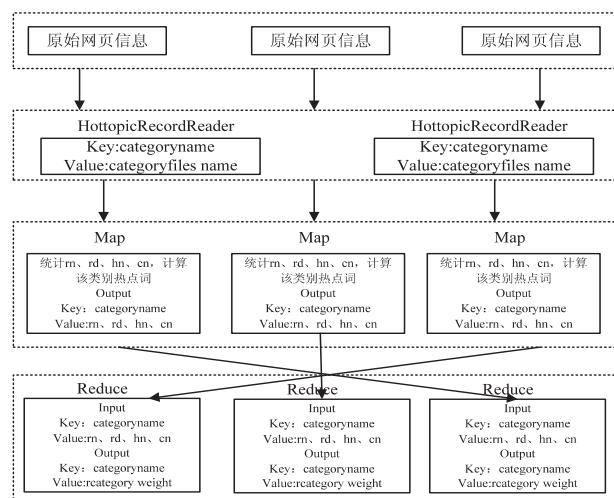


图3 热点计算流程

map 函数有两部分功能。第一部分是计算热点(计算上述1,4,6,7项,2,3,5项不能计算是因为需要其他类别的数据)。将计算的1,4,6,7项传到 reduce 函数计算剩下的2,3,5项,并最终计算热点。

### 3.2.4 话题跟踪

文中采用基于查询的话题跟踪<sup>[17]</sup>方法。系统定

义两个阈值:跟踪阈值  $t_1$  和跟踪器调整阈值  $t_2$ 。其中,  $t_1 < t_2$ 。

话题查询向量的构建过程如下:将训练集中出现的非停用词  $w$  按照其对应的  $r * \text{idf}(w)$  值由高到低排序,其中  $r$  为包含词  $w$  的相关报道数量,  $\text{idf}(w)$  为词  $w$  的  $\text{idf}$  值;取前  $n$  个词组成查询向量。查询向量第  $k$  维(词  $w$ )的取值  $q(k) = \text{tf}(w) * \text{idf}(w)$ ,  $\text{tf}(w)$  为所有相关报道中词  $w$  的平均  $\text{tf}$  值。对应跟踪器中第  $k$  维(词  $w$ ),新闻报道向量的第  $k$  维的取值:

$$d(k) = 0.4 + 0.6 * \text{tf}(w) * \text{idf}(w)$$

新闻报道  $d$  和跟踪器  $q$  的相似度值采用加权和方法计算,其中  $q_k$  和  $d_k$  分别表示各自向量中第  $k$  个词的权重。

$$\text{sim}(q, d) = \frac{\sum_{k=1}^N q_k * d_k}{\sum_{k=1}^N q_k} \quad (2)$$

算法描述如下:

```
if ( sim(q, d) > t1 )
{ 判定 d 为相关报道
if ( sim(q, d) > t2 )
{ 重构跟踪器 q 以吸收该话题重要的新特征; }
```

跟踪器  $q$  的重构过程:

(1) 确定查询向量  $q$  的核心特征项,令权重最大的5个特征项为核心特征项,核心特征项在最初建立查询向量时确定,不参与任何调整。

(2) 判断权重最大的文本向量特征项的权重  $w_d$  与权重最大的查询向量非核心特征项  $w_q$  的大小,如果  $w_d > w_q$  则转到第3步,否则转到第4步。

(3) 用该文本向量权重最大的特征项替换查询向量权重最小的非核心特征项,替换后的特征权重为向量权重与相似度的积,即  $w_d * \text{sim}(q, d)$ 。

(4) 结束对查询向量的调整。

## 4 实验结果分析

### 4.1 系统运行环境

系统采用8台HP商用服务器,服务器操作系统为64位CentOS6.4, Hadoop采用hadoop1.2.1版本, Java采用64位jdk1.7.0\_60版本。采用Hadoop集群Namenode一台, SecondNamenode一台,其余作为Data-node。

### 4.2 评测机制

文中依据TDT评测标准,采用漏报率(Miss)、误报率(False Alarm, FA)以及识别代价(CDet) Norm来评价话题聚类的性能,话题  $i(i = 1, 2, \dots, k)$  的漏报率和误报率定义为:

$$Miss_i = \frac{\text{未检测到的与话题相关的报道数}}{\text{与话题 } i \text{ 相关的报道数}}$$
$$FA_i = \frac{\text{检测到的与话题不相关的报道数}}{\text{与话题 } i \text{ 不相关的报道数}} \quad (3)$$
$$(C_{Det})_{Norm} = \frac{C_{Miss} * P_{Miss} * P_{target} + C_{Fa} * P_{Fa} * P_{\neg target}}{\min(C_{Miss} * P_{target}, C_{Fa} * P_{\neg target})}$$

其中,  $C_{Miss}$  和  $C_{Fa}$  分别是漏报和误报的代价;  $P_{Miss}$  和  $P_{Fa}$  分别是漏报和误报的条件概率;  $P_{target}$  是目标话题的先验概率;  $P_{\neg target} = 1 - P_{target}$ 。

$C_{Miss}$ 、 $C_{Fa}$  和  $P_{target}$  都是预设值,用来调节漏报率和误报率在评测结果中所占比重。评测中三个参数分别取 1.0,0.1,0.02。(CDet)Norm 是系统的性能评测指标,该值越小,表明算法的性能越好。

4.3 实验结果

实验 1:  $K$ -means 与 Canopy-Kmeans 实验对比。  
该实验采用系统网络爬虫抓取的 6 573 条数据作为信息来源,实验的目的是通过漏报率、误报率和识别代价对传统  $K$ -means 和文中所使用的 Canopy-Kmeans 进行对比分析,结果见表 1。

表 1 聚类实验对比 %

| 话 题     | K-means |      |       | 话 题     | Canopy-Kmeans |      |       |
|---------|---------|------|-------|---------|---------------|------|-------|
|         | Miss    | FA   | Cdet  |         | Miss          | FA   | Cdet  |
| 海洋仪器分析  | 36.79   | 4.03 | 56.54 | 科研经费乱象  | 39.14         | 3.97 | 58.59 |
| 天力业绩锐减  | 38.63   | 3.98 | 58.13 | 云计算政务应用 | 24.12         | 3.96 | 53.72 |
| 二次装修浪费  | 37.15   | 3.52 | 54.39 | 公车改革补贴  | 36.25         | 4.21 | 56.87 |
| 资源不均致房热 | 40.13   | 4.02 | 56.57 | 二次装修浪费  | 36.13         | 3.01 | 50.88 |
| ...     | ...     | ...  | ...   | ...     | ...           | ...  | ...   |
| 土地限购    | 39.36   | 3.74 | 57.69 | 天力业绩锐减  | 37.29         | 3.56 | 54.73 |
| 平均值     | 38.49   | 3.84 | 57.31 | 平均值     | 37.25         | 3.75 | 55.63 |

通过实验结果可以看出,文中使用的聚类方法漏报率平均降低 1.24%,误报率平均降低 0.09%,最小标准代价平均降低 1.681%。

实验 2:单机与分布式时间开销比较。  
该系统采用网络爬虫抓取的 6 573 条数据作为信息来源,从中抽取 1 000,2 500,4 500 和 6 573 条分别采用单机和分布式模式进行聚类以及热点计算,结果见图 4。

由实验数据可以看出,当系统中的文档数目较少时,传统的单机模式比 MapReduce 方式更快,但是随着文档规模的扩大,MapReduce 方式便体现出了分布式计算的优点。

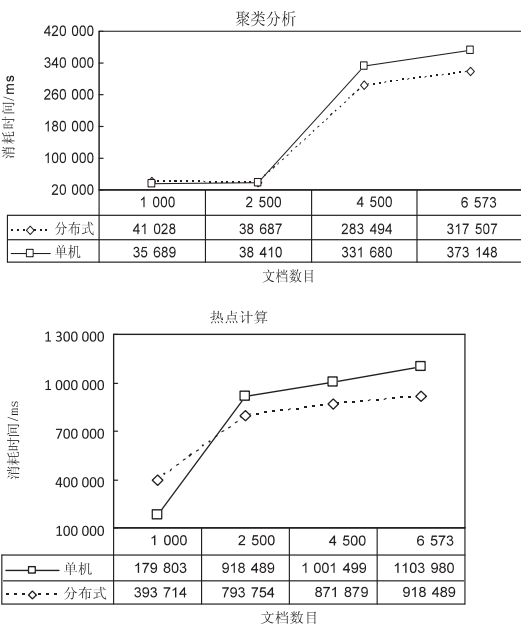


图 4 时间开销比较

实验 3:热点话题展示。

通过对爬虫抓取到的 6 573 条记录进行聚类分析和热点计算,得分前五的话题,见表 2。

表 2 热点话题发现及实验结果 %

| 排名  | 得分   | 报道数 | 关键词                   | Miss  | FA   | Cdet  |
|-----|------|-----|-----------------------|-------|------|-------|
| 1   | 7.20 | 760 | 山东, 科学院, 科技, 发展...    | 34.13 | 3.12 | 49.42 |
| 2   | 6.94 | 710 | 山东, 海洋仪器, 发展, 分析...   | 33.25 | 3.56 | 50.69 |
| 3   | 6.51 | 650 | 山东, 产业, 模型, 促进, 发展... | 35.26 | 3.27 | 51.28 |
| 4   | 5.07 | 500 | 山东, 天力公司, 业绩, 锐减...   | 32.17 | 3.01 | 46.92 |
| 5   | 4.95 | 430 | 济南, 莱芜, 合作, 促进, 发展... | 31.98 | 3.29 | 48.10 |
| 平 均 |      |     |                       | 33.36 | 3.25 | 49.28 |

通过与抓取的新闻报道进行对比,该系统采用的热点发现算法得出的实验结果与报道情况符合。从实验结果看出得分前五的热点平均漏报率为 0.333 6,平均误报率为 0.032 5,最小标准代价为 0.492 8,验证了系统所使用的热点计算方法有效可行。

5 结束语

文中设计并实现了一个基于 Hadoop 的网络舆情系统。该系统利用开源的 Mahout 工具以及 MapReduce 编程模式来实现基于朴素贝叶斯的文本分类算法,基于 Canopy-Kmeans 的文本聚类算法,综合考虑新闻报道数量、点击量、回复量、报道来源等因素的热点计算和基于查询向量的话题跟踪算法等。

今后的研究工作中,将继续对系统进行完善,如:增加分布式实时索引和检索功能,采用数字签名技术

对网页信息去重,进一步丰富舆情信息展示等。

参考文献:

[1] 陈彦舟,曹金璇. 基于Hadoop的微博舆情监控系统[J]. 计算机系统应用,2013,22(4):18-22.

[2] 王宏宇. Hadoop平台在云计算中的应用[J]. 软件,2011,32(4):36-38.

[3] Owen S, Anil R, Dunning T, et al. Mahout in action[M]. [s. l.]: Manning Publications, 2011.

[4] McCallum A, Nigam K, Ungar L H. Efficient clustering of high-dimensional data sets with application to reference matching[C]//Proc of the 6th ACM SIGKDD. [s. l.]: ACM, 2000: 169-178.

[5] MacQueen J. Some methods for classification and analysis of multivariate observations[C]//Proc of the 5th Berkeley symposium on mathematical statistics and probability. California: University of California Press, 1967:281-287.

[6] 董坚峰. 面向公共危机预警的网络舆情分析研究[D]. 武汉:武汉大学,2013.

[7] 王宇阳. 基于本体进化的自适应中文话题跟踪算法研究[D]. 南京:南京航空航天大学,2013.

[8] Schultz J, Liberma M. Topic detection and tracking using IDF-weighted cosine coefficient[C]//Proceedings of the DARPA

broadcast news workshop. Herndon: [s. n.], 1999:189-192.

[9] 龚海军. 网络热点话题自动发现技术研究[D]. 武汉:华中师范大学,2008.

[10] Shvachko K, Kuang H, Radia S, et al. The Hadoop distributed file system[C]//Proc of IEEE 26th symposium on mass storage systems and technologies. [s. l.]: IEEE, 2010.

[11] Dean J, Ghemawat S. MapReduce simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1):107-113.

[12] Vashishtha H, Stroulia E. Enhancing query support in HBase via an extended coprocessors framework[C]//Proceedings of the European conference on towards a service-based internet. [s. l.]: [s. n.], 2011:75-87.

[13] 李猛. 基于DOM的Web信息抽取技术的研究与实现[D]. 大连:大连理工大学,2008.

[14] 钱浩. Web信息抽取技术的研究与应用[D]. 大庆:东北石油大学,2011.

[15] 王星. 新闻网页抽取技术的研究与实现[D]. 天津:河北工业大学,2011.

[16] 高岩. 朴素贝叶斯分类器的改进研究[D]. 广州:华南理工大学,2011.

[17] 邹鸿程. 微博话题检测与追踪技术研究[D]. 郑州:解放军信息工程大学,2012.

(上接第143页)

[5] 林飞跃,林先津. 云桌面在教学管理中的应用[J]. 实验室研究与探索,2013,32(10):336-339.

[6] 虚拟化与云计算小组. 云计算宝典:技术与实践[M]. 北京:电子工业出版社,2011.

[7] 桌面虚拟化[EB/OL]. 2015. [http://baike.baidu.com/link?Url=nsaINMEbnJ\\_bid5tuEdkS306X09k1VLTpdW\\_WI6\\_rEq-VNvwsUuaJy1q0-95JPCrxUOEucPX2KspSejJhB\\_GGa](http://baike.baidu.com/link?Url=nsaINMEbnJ_bid5tuEdkS306X09k1VLTpdW_WI6_rEq-VNvwsUuaJy1q0-95JPCrxUOEucPX2KspSejJhB_GGa).

[8] 刘剑锋. 浅谈虚拟化桌面在高校的架构和应用[J]. 网络安全技术与应用,2012(11):71-72.

[9] 李力. 基于桌面虚拟化的高校机房设计和建设[J]. 价值工程,2014,33(8):219-220.

[10] 桌面虚拟化的几大维度[EB/OL]. 2015. <http://soft.chinabyte.com/120/12860620.shtml>.

[11] H3C CAS VDI 虚拟桌面产品技术白皮书[EB/OL]. 2015. [http://wenku.baidu.com/link?url=HiELhi99Z9VYolymBVfMX-9bS-XSMFw4X\\_OlUQrmmH\\_2l1jKAfuKqUbIM6qdyyMdc5HVLcqZzuyGAXotGsa-TieD1Xj-mQjV\\_XKLnHT272W](http://wenku.baidu.com/link?url=HiELhi99Z9VYolymBVfMX-9bS-XSMFw4X_OlUQrmmH_2l1jKAfuKqUbIM6qdyyMdc5HVLcqZzuyGAXotGsa-TieD1Xj-mQjV_XKLnHT272W).

[12] 和信窗体科技有限公司. VEMS系统2.0产品白皮书[EB/

OL]. 2015. [http://Wenku.baidu.com/link?url=PbrjMS9kYtDYCUnnXHVcBBEPTFLpGDT0rycLs1U1As3zakThux\\_ea1sLyAsLpGMEZ6Sk85LXoXFxrWCs3Q7y4ahzml2sBA0vIqPZAL8eM3](http://Wenku.baidu.com/link?url=PbrjMS9kYtDYCUnnXHVcBBEPTFLpGDT0rycLs1U1As3zakThux_ea1sLyAsLpGMEZ6Sk85LXoXFxrWCs3Q7y4ahzml2sBA0vIqPZAL8eM3).

[13] 白伟,李凤英. 浅谈桌面虚拟化技术发展与应用现状[J]. 中小企业管理与科技,2013(34):280-282.

[14] Windows\_Server\_2012\_R2\_VDI\_Datasheet[EB/OL]. 2015. <http://www.microsoft.com/en-us/server-cloud/products/virtual-desktop-infrastructure/Overview.aspx>.

[15] 黄金敢. 高校教学环境中桌面云架构研究与实现[J]. 计算机技术与发展,2013,23(12):222-225.

[16] 项国富,金海,邹德清,等. 基于虚拟化的安全监控[J]. 软件学报,2012,23(8):2173-2187.

[17] 何永忠,王伟,黎琳. 基于云计算的信息安全实验教学平台建设[J]. 计算机教育,2014(1):39-42.


[18] 周相兵,余堃,马洪江. 一种云服务的质量模型研究[J]. 小型微型计算机系统,2013,34(12):2718-2723.

[19] Lombardi F, Pietro R D. Secure virtualization for cloud computing[J]. Journal of Network and Computer Applications, 2011,34:1113-1122.

基于Hadoop的网络舆情监控平台设计与实现

作者：[李晨](#)，[杨子江](#)，[朱世伟](#)，[于俊凤](#)，[LI Chen](#)，[YANG Zi-jiang](#)，[ZHU Shi-wei](#)，[YU Jun-feng](#)

作者单位：[山东省科学院 情报研究所, 山东 济南, 250014](#)

刊名：[计算机技术与发展](#)

英文刊名：

年，卷(期)：2016 (2)

引用本文格式：[李晨](#).[杨子江](#).[朱世伟](#).[于俊凤](#).[LI Chen](#).[YANG Zi-jiang](#).[ZHU Shi-wei](#).[YU Jun-feng](#) [基于Hadoop的网络舆情监控平台设计与实现](#) [期刊论文]-[计算机技术与发展](#) 2016 (2)