基于 Hadoop 的微博舆情监控系统①

陈彦舟, 曹金璇

(中国人民公安大学 信息安全工程系, 北京 102600)

摘 要: 随着在线社会网络如社交网站、微博、在线社区等的快速发展, 一个真正的双向传播和新媒体时代逐步形成. 在线社会网络让每个用户都能创造自己的内容, 并且快速传播出去. 据不完全统计, 新浪微博平均每秒有超过 1000 条的新微博产生, 日增量数据为 5TB, 因此海量数据给舆情监控带来了严峻的挑战. 将介绍一种基于 Hadoop 的微博舆情监控系统, 能够对大规模采集数据进行挖掘、分析, 实现对舆情热点话题的发现及追踪、对微博的社会网络分析, 分析结果可视化呈现, 为党政机关、大型企业等单位和组织及时发现敏感信息、掌握舆情热点、把握舆情趋势、应对舆论危机提供自动化、系统化、科学化的信息支持.

关键词: 舆情监控; Hadoop; HBase; MapReduce; 在线社会网络; 云计算

Public Sentiment Monitoring System for Microblog Based on Hadoop

CHEN Yan-Zhou, CAO Jin-Xuan

(Department of Information Security Engineering, Chinese People's Public Security University, Beijing 102600, China)

Abstract: With the rapid development of online social networks, such as social networking services, microblog, online community, etc., a real two-way communication and new media age has been gradually forming. Everyone can create their own content and spread out quickly through online social networks. According to incomplete statistics, Sina microblog generates over 1000 new microblogs per second and the daily incremental size of data is 5TB. Thus, massive data has brought severe challenge to public opinion monitoring. This article will introduce a microblog public opinion monitoring system based on the Hadoop. It can mine and analyze large scale collected data, realize detection and tracking of hot topics, perform social network analysis on the microblog and visualize the analysis result. The proposed system will provide automated, systematic, and scientific information support for party and government organizations, enterprises and other units and organizations to detect sensitive information timely, grasp the hot points and the trend of public opinion and deal with the crisis of public opinion.

Key words: monitoring public opinion; Hadoop; HBase; MapReduce; online social networks (OSN); cloud computing

舆情[1]是指在一定的社会空间内, 围绕中介性社会事件的发生、发展和变化, 它是较多群众关于社会中各种现象、问题所表达的信念、态度、意见和情绪等等表现的总和. 互联网开放、虚拟的特性让言论达到了前所未有的活跃程度. 对于公众关注的事件很快就会形成网上舆论, 个别人在其中煽风点火, 极易造成网络非理性情绪蔓延, 并产生严重的不良影响, 对相关部门造成巨大的舆论压力.

随着互联网的快速发展, 世界范围内的互联网用户也在急剧膨胀, 同时出现了一批优秀的在线社会网络, 而其中最为活跃的是社交网站和微博领域, 比如: 美国的 FaceBook、Twitter; 俄罗斯的 Odnoklassniki、Vkontakte; 越南的 Zing; 中国的校内网、Qzone、腾讯微博以及新浪微博等等. 据不完全统计, 腾讯微博目前

① 基金项目:中央高校基本科研业务费专项资金(YX11133)
收稿时间:2012-09-21;收到修改稿时间:2012-11-07

注册用户达 4.25 亿, 平均每秒产生 50 条新消息, 人均好友量 120 个, 每月产生 19 亿消息量. 新浪微博在不到三年的时间已积累了近 3 亿用户, 平均每秒有超过 1000 条的新微博产生[2], 日增量数据为 5TB. 随着海量的微博消息不断地被创造出来, 如何从这些海量的数据中进行挖掘、分析, 实现对敏感信息及舆情热点话题的持续追踪及传播趋势研判成为一个重要研究方向以及挑战.

传统的舆情监控系统都是基于昂贵的工作站或服务器集群, 在面对海量数据时往往表现为成本高昂、可扩展性差、单点通信故障等等, 再加上传统数据库难以管理和批量处理上亿数据记录, 因此我们开发了一个基于Hadoop的微博舆情监控系统, 能够对大规模采集数据进行挖掘、分析, 实现对舆情热点话题的发现及追踪、对微博的社会网络分析, 分析结果可视化呈现, 为党政机关、大型企业等单位和组织及时发现敏感信息、掌握舆情热点、把握舆情趋势、应对舆论危机提供自动化、系统化、科学化的信息支持.

本文主要介绍以下方面的内容: 第一部分介绍微博舆情监控系统框架. 第二部分介绍微博舆情监控系统结构. 第三部分介绍系统实现.

1 系统框架

Hadoop[3]是 Apache 开源分布式系统基础架构, 因其优良的特性而被大量运用于企业和机构的研究, 以及被用来搭建自己的云计算平台. Hadoop 主要是由 HDFS(Google File System(GFS) 的开源实现)、MapReduce[4](Google MapReduce 的开源实现)和 HBase[13](Google BigTable 的开源实现)组成. 微博舆情监控系统基于 Hadoop, 以 HBase 做为海量采集数据存储数据库, 实现微博信息采集、微博舆情监控以及用户交互三层. 如图 1 所示为微博舆情监控系统的系统架构, 其中微博信息采集和清洗过滤层完成对微博数据的采集和清洗, 分布式计算层完成热点话题和社交网络分析, 分布式存储层用于存储采集的数据以及分析结果.

2 微博舆情监控系统结构

2.1 微博舆情监控系统总体结构

微博舆情监控系统由微博信息采集模块、微博信息分析模块、索引存储模块、舆情监控分析模块、交互模块组成. 每个部分的功能如下:

(1) 微博信息采集模块: 采集微博博主信息和微博内容;

(2) 微博信息分析模块: 信息抽取、网页消重、文本切词;

(3) 索引存储模块: 提供对 Hadoop 分布式数据(索引库、HBase 库、分析库)的操作接口;

(4) 舆情监控分析模块: 文本表示、对索引库和 HBase 库里的数据进行聚类分析、社会网络分析等, 结果输出到分析库;

(5) 交互模块: 基于 J2EE 架构实现用户交互功能.

如图 2 所示为微博舆情监控系统结构与功能.

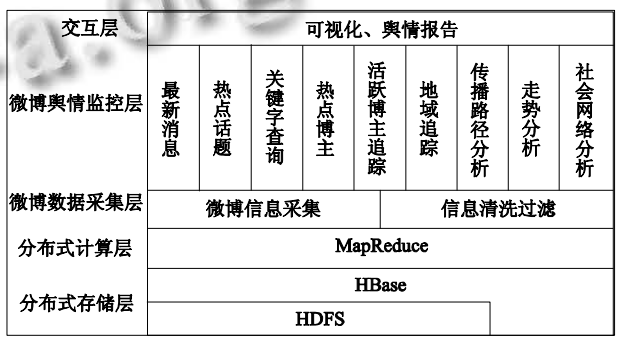

图 1 微博舆情监控系统架构

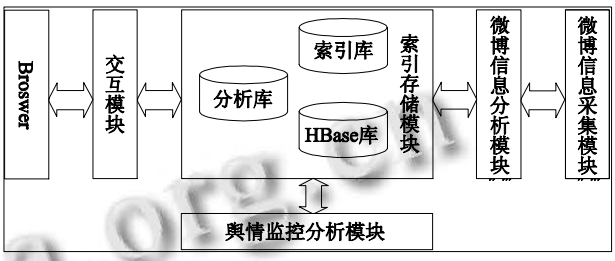

图 2 微博舆情监控系统结构与功能

2.2 微博信息采集及分析

微博信息采集采用文献[5]提出的基于 API 与网页解析方案相结合的方法. Open API 是指一种微博服务商将自己提供的服务封装成一系列 API 接口, 通过调用这些数据接口可以获取微博内容、评论、用户、关系等信息. 其中新浪和腾讯提供的 API 最为丰富, 而且新浪微博是国内最大的在线社会网络, 因此本文信息采集的来源定为新浪微博. 为了均衡服务器的负载, 微博服务商对不同用户设置了不同的 API 接口调用频率与查询范围. 新浪微博不仅限制了一次请求最多只能返回 5000 个结果和普通授权用户每小时接口最多只能使用 1000 次, 而且拒绝短时间内高频率的 API 接

口调用. 因此在采集微博信息中我们采用了基于 API 与网页解析方案相结合的方法, 如图 3 所示.

(1) 获取器: 通过调用 API 接口返回 JSON 格式文件方式收集博主信息;

(2) 爬虫器: 通过分布式爬虫方法抓取微博内容, 并利用 Dom 解析 html 和抽取信息.

其中 n 个获取器和爬虫器分别运行在 n 个 slaver 机器上, 调度器运行在 master 机器上.

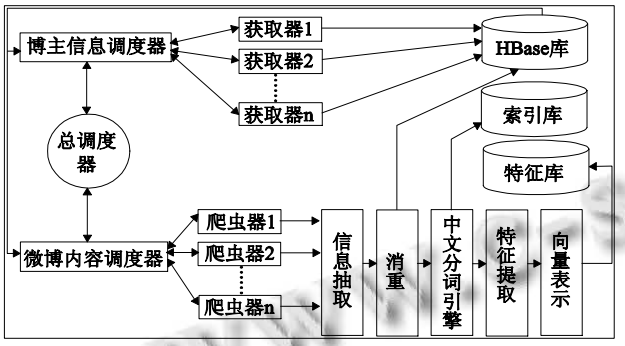

图 3　微博信息采集及分析模块结构

本系统利用词频-逆向文本频率 TFIDF 向量表示法来表示微博内容的特征向量. 在不考虑词间次序和文本结构的前提下, 这种表示方法针对微博内容中的所有词(在文本切词阶段对去除@微博用户和短链接的微博内容, 通过庖丁解牛分词处理取得词语), 所以本质上讲它也是一种微博内容的词集表示法. 类似于结构化数据库的一条记录, 一条微博内容的 TFIDF 特征向量某种程度上反映了该微博的内容特征. 以一个矩阵来表示所有微博内容集合文本信息, 矩阵中的列集为特征集, 行集为所有已爬取的微博内容集合. 如图 3 所示, 微博内容写入 HBase 库, 微博索引写入索引库, 特征矩阵写入特征库.

2.3 舆情监控分析

舆情监控分析作为系统的主要模块, 它包含了最新消息、热点话题、热点博主、活跃博主追踪、地域追踪、传播路径分析、走势分析、社会网络分析. 限于篇幅, 以下我们仅对主要功能展开具体的阐述.

2.3.1 热点话题

热点话题是在上述构造出的特征矩阵基础上做聚类分析, 整个并行计算流程如图 4 所示读取特征库中的特征向量, 利用 Canopy 算法[7]对 K-Means 聚类算法[8]优化的方法计算出相似内容, 最后把各中心点以及包含的子项写入分析库供前台查询. 工作流程如下:

(1) 读取特征库中的特征矩阵;

(2) 通过基于 MapReduce 的 Canopy 算法取得中心;

(3) 通过基于 MapReduce 的 K-Means 算法计算数据对象与聚类中心间距离;

(4) 把聚类结果中各中心点以及包含的子项写入分析库供前台查询.

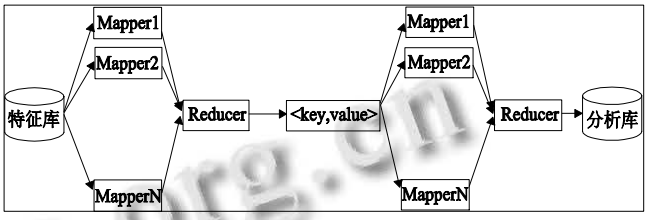

图 4　热点话题发现模块结构

2.3.2 社会网络分析

复杂网络[9]是近几年在计算机处理和运算能力飞速发展基础上被人们发现的真实网络. 它具有很多与规则网络和随机网络不同的统计特征, 其中最重要三个特性就是小世界特性、无标度特性以及高聚类系数. 社会网络的概念与复杂网络相比, 二者之间有很多共同的地方, 首先二者都是通过网络(G=(V,E))的概念来描述被研究的对象及对象之间的关系, 其次研究复杂网络的主要方法中包括社会网络分析方法. 社会网络和复杂网络之间的关系是被包容与包容的关系, 社会网络是复杂网络的一种, 是复杂网络研究领域中一种特殊的网络[10]. 本文通过节点度分布和聚类系数来分析微博网络无标度性和小世界特性.

在微博网络中每个节点表示一个博主. 对于一个博主, 与其有连接的其它博主的个数称为该博主的度. 一个博主的入度指的是其粉丝总和, 其出度指的是关注总和. 在某种意义上, 节点的度数越高, 说明该博主越重要. 因此, 研究分析整个社会网络中的节点度分布的情况, 可以帮助了解社会网络的结构.

聚类系数是社会网络分析中的一个重要的指标, 它是指社会网络中实际存在的边数和可能有的边数之比. 聚类系数反应了网络的集团化程度, 这是一种网络内聚的反映, 对于社会网络而言, 集团化形态是一个重要特征, 集团表示网络中的朋友圈或熟人圈的凝聚力的程度, 集团中的成员往往相互熟悉, 聚类系数就是刻画这种群集现象的集团化属性[10].

本系统所采用的社交网络分析法是计算博主信息中的粉丝数以及关注数的出入度和聚类系数, 分别采

用 n 个 Map 阶段和一个 Reduce 阶段, 计算结果存储在分析库, 供前台可视化交互调用.

3 系统实现

3.1 集群系统结构

由连接在千兆以太网交换机上的 1 个 Master (ubmaster)及 3 个 Slaves(ubslave1、ubslave2、ubslave3)构成, IP 分别为 192.168.102.230～192.168.102.233, 其中 Master 运行 NameNode 和 JobTracker, 子节点运行 DataNode 及 TaskTracker, 所有机器均为方正台式机, 每台内存 2G, 空间 250G, 软件环境详细信息如表 1 所示, 节点拓扑结构图如图 5 所示.

表 1 各节点硬件及软件配置

硬件		软件		
CPU	Intel(R)Core(TM)2 E8400 3.00GHz×2	操作系统	Ubuntu 10.04 LTS	
内存	4GB	其他	Master	Slaves
硬盘	498GB WD5000AAKS 7200RPM	Java JDK7	√	√
以太网卡	RTL8168D(P)/8111D(P) PCI-E Gigabit Ethernet	SSH	√	√
交换机	1000Mbps	Hadoop-0.20.2	√	√
		HBase-0.90.4	√	√
		Eclipse3.7.1	√	
		Hive-0.8.1	√	
		Tomcat-7.0.22	√	
		Lucene-3.0.3	√	

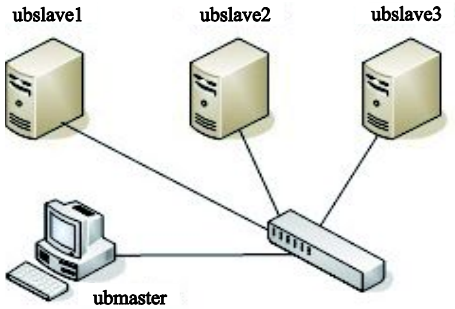

图 5 节点拓扑结构图

3.2 环境配置

3.2.1 SSH 配置

① 在 ubmaster 机器上建立 SSH Key: ssh-keygen －t rsa －P “”

② 启动 SSH Key: cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys sudo /etc/init.d/ssh reload

③ 在 ubmaster 机器上向 3 个 slave 机器上的 /home/hadoop/.ssh/文件夹里发放公钥 authorized_keys

④ 测试无密码连接: ssh ubslave1、ssh ubslave2、ssh ubslave3

3.2.2 Hadoop、HBase、Hive 环境配置

在启动 Hadoop、HBase、Hive 环境前需要配置 Hadoop 基本环境配置文件 hadoop-env.sh、core-site.xml、mapred-site.xml 、 hdfs-site.xml 、 hbase-env.sh 、 regionservers、hbase-site.xml、hive-env.sh、hive-site.xml, 有关这些文件的配置参数, 见文献[11], 这里不再赘述.

3.3 HBase 数据库设计

HBase 数据库存储采集的博主信息、微博内容以及关注信息, 如表 2 所示.

表 2 微博舆情监控系统数据库设计

Rowkey	TimeStamp	Column Family	Column Keys
id	timestamp	user	id,sex,brithday,address,gzNum,fansNum,wbNum,edu,work,summary,blog,renZh
		weibo	id,person_id,article,discuss,transmit,origin,time
		userR	idA, idB

其中 user 列簇存储博主信息, weibo 列簇存储微博内容, userR 存储博主关系(博主 idA 关注博主 idB).

3.4 系统实现

限于篇幅, 以下我们仅对主要实现功能展开具体的阐述.

3.4.1 微博信息采集

在系统实现阶段, 我们总共采集了将近 100 万博主信息和 22 万微博内容. 图 6 所示为采集进 HBase 库中的局部微博内容, 其中 article 字段是中文以十六进制存储.

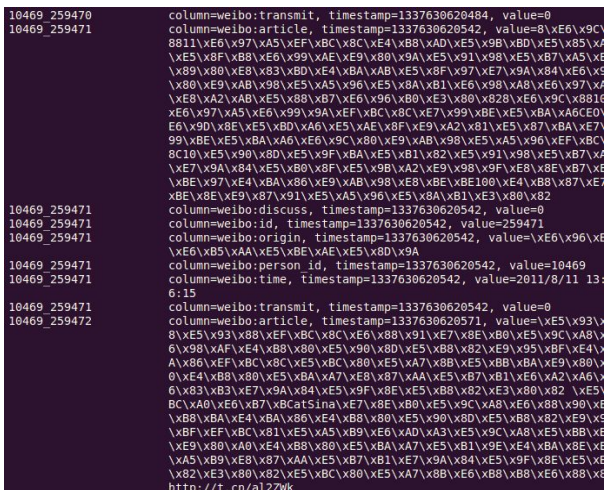
10469_259470 column=weibo:transmit, timestamp=1337630620484, value=0
10469_259471 column=weibo:article, timestamp=1337630620542, value=8\xE6\x9C\
8811\xE6\x97\xA5\xEF\xBC\x8C\xE4\xB8\xAD\xE5\x9B\xBD\xE5\x85\x
\xE5\x8F\x88\xE6\x99\xAE\xE9\x80\x9A\xE5\x91\x98\xE5\xB7\xA5\x
\x89\x80\xE8\x83\xBD\xE4\xBA\xAB\xE5\x8F\x97\xE7\x9A\x84\xE6\x
\x80\xE9\xAB\x98\xE5\xA5\x96\xE5\x8A\xB1\xE6\x98\xA8\xE6\x97\x
\xE8\xA2\xAB\xE5\x88\xB7\xE6\x96\x80\xE3\x80\x828\xE6\x9C\xB81
xE6\x97\xA5\xE6\x99\x9A\xEF\xBC\x8C\xE7\x99\xBE\xE5\xBA\xA6CEO\
E6\x9D\x8E\xE5\xBD\xA6\xE5\xAE\x8F\xE9\xA2\x81\xE5\x87\xBA\xE7\
99\xBE\xE5\xBA\xA6\xE6\x9C\x80\xE9\xAB\x98\xE5\xA5\x96\xEF\xBC\
8C10\xE5\x90\x8D\xE5\x9F\xBA\xE5\xB1\x82\xE5\x91\x98\xE5\xB7\x
\xE7\x9A\x84\xE5\x80\x8F\xE5\x9B\xA2\xE9\x98\x9F\xE8\x8E\xB7\x
\xBE\x97\xE4\xBA\x86\xE9\xAB\x98\xE8\xBE\xBE100\xE4\xB8\x87\xE7
xBE\x8E\xE9\x87\x91\xE5\xA5\x96\xE5\x8A\xB1\xE3\x80\x82
10469_259471 column=weibo:discuss, timestamp=1337630620542, value=0
10469_259471 column=weibo:id, timestamp=1337630620542, value=259471
10469_259471 column=weibo:origin, timestamp=1337630620542, value=\xE6\x96\x
\xE6\xB5\xAA\xE5\xBE\xAE\xE5\x8D\x9A
10469_259471 column=weibo:person_id, timestamp=1337630620542, value=10469
10469_259471 column=weibo:time, timestamp=1337630620542, value=2011/8/11 13:
6:15
10469_259471 column=weibo:transmit, timestamp=1337630620542, value=0
10469_259472 column=weibo:article, timestamp=1337630620571, value=\xE5\x93\x
8\xE5\x93\x88\xEF\xBC\x8C\xE6\x88\x91\xE7\x8E\xB0\xE5\x9C\xA8\
6\x98\xAF\xE4\xB8\x80\xE5\x90\x8D\xE5\xB8\x82\xE9\x95\xBF\xE4\
A\x86\xEF\xBC\x8C\xE5\xBC\x80\xE5\xA7\x8B\xE5\xBB\xBA\xE9\x80\
0\xE4\xB8\x80\xE5\xBA\xA7\xE8\x87\xAA\xE5\xB7\xB1\xE6\xA2\xA6\
6\x83\xB3\xE7\x9A\x84\xE5\x9F\x8E\xE5\xB8\x82\xE3\x80\x82 \xE5\
BC\xA0\xE6\xB7\xBCatSina\xE7\x8E\xB0\xE5\x9C\xA8\xE6\x88\x90\x
\xB8\xBA\xE4\xBA\x86\xE4\xB8\x80\xE5\x90\x8D\xE5\xB8\x82\xE9\x
\xBF\xEF\xBC\x81\xE5\xA5\xB9\xE6\xAD\xA3\xE5\x9C\xA8\xE5\xBB\x
\xE9\x80\xA0\xE4\xB8\x80\xE5\xBA\xA7\xE5\xB1\x9E\xE4\xBA\x8E\x
\xA5\xB9\xE8\x87\xAA\xE5\xB7\xB1\xE7\x9A\x84\xE5\x9F\x8E\xE5\x
\x82\xE3\x80\x82\xE5\xBC\x80\xE5\xA7\x8B\xE6\xB8\xB8\xE6\x88\x
http://t.cn/zl2Zwk

图 6 微博采集图

3.4.2 热点话题发现及可视化

用户通过交互模块读取分析库中的聚类中心以及其包含的子项，然后使用力导向算法和JavaScript进行可视化，如图 7 所示：中心点为当日的话题，其周围的小点代表参与此话题的博主.

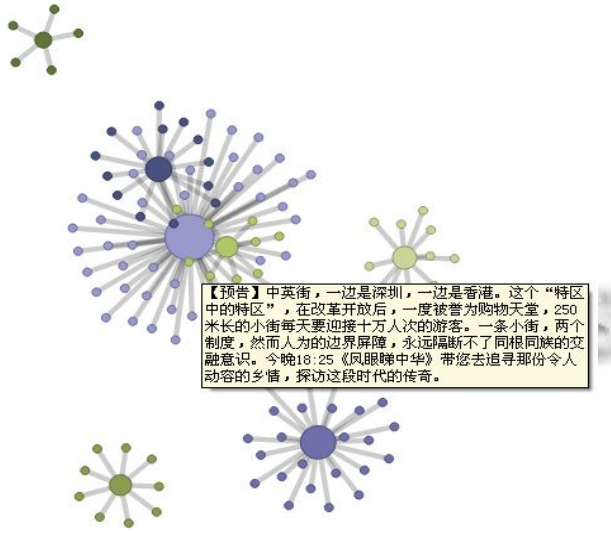

图 7　热点话题可视化图

3.4.3 社会网络分析

复杂网络最重要三个特性就是小世界特性、无标度特性以及高聚类系数，本系统通过节点度分布和聚类系数来分析微博网络无标度性和小世界特性. 通过对博主信息中的关注和粉丝关系进行 MapReduce 计算后，结果如图 8 所示，其中上图为节点出度分布图、下图为节点入度分布图. 其中由图 8 中上图可知在出度为 160 左右会出现一个尖峰，由统计数据可知出度在 150 到 160 之间的用户为 92 人. 由下图可知有 200 以上人的跟随者大于 1000 人，其中有 62 人发表的微博在 100 条以上，所以这部分用户是活跃用户，他们不但拥有数量众多的好友，而且也有不少跟随者，并且经常发表微博，这些用户是网络中活动较强的群体.

4　结语

本文主要对分布式系统关键技术进行了研究，并把 Hadoop 分布式存储和 MapReduce 并行计算模型运用于海量微博数据处理当中，对微博进行了舆情监控分析. 本文主要完成了以下工作:

(1) 利用基于 API 与网页解析方案相结合的方法从微博获取博主信息和爬取微博内容;

(2) 结合 HBase 的架构特点和 MapReduce 并行计算方法对舆情监控研究分析，并设计构建出基于 HBase 的微博信息存储系统，基于 MapReduce 的分布式热点话题算法;

(3) 通过多种可视化方法有效展示了舆情监控分析结果.

通过实践证明，基于Hadoop的微博舆情监控系统可以有效的对大规模微博数据进行舆情监控分析，下一步我们的主要工作:

(1) 使用 CTM 模型改进话题聚类算法;

(2) 尝试基于模型的多维聚类算法，提高获取热点话题的准确度;

(3) 进一步结合社会网络分析理论，增强舆情监控准确度.

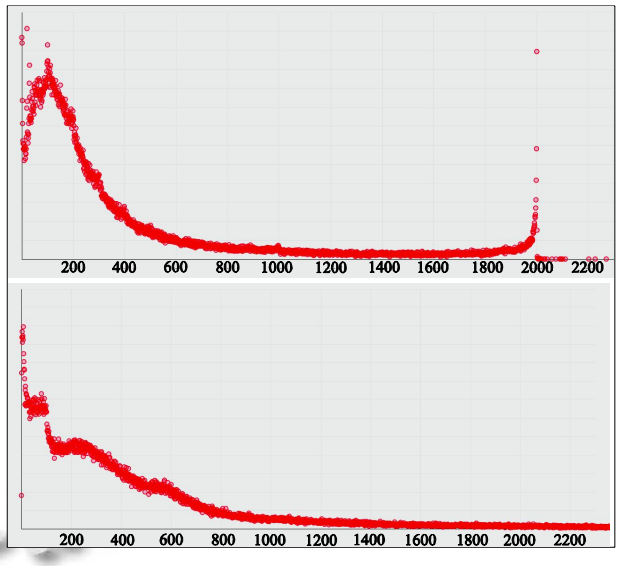

图 8　社会网络分析可视化图

参考文献

1 刘毅.略论网络舆情的概念、特点、表达与传播.前沿论坛, 2007,(l):11–12.

2 Zheng J. “社交网络世界地图”最新版发布. [2012-06-11]. http://www.36kr.com/p/117083.html.

3 Apache Hadoop.Hadoop.[2011202215] http://hadoop.apache.org/

4 Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Operating Systems Design and Implementation, 2004. 137–149.

5 廉捷,周欣,曹伟,刘云.新浪微博数据挖掘方案.清华大学学

(下转第 9 页)

③ 调用方法表中的方法并依据知识结构进行推理. 方法可以分为两种类型, 即“机器提问人为回答”方式或“机器提问机器回答”方式. 推理机根据通过解析的知识结构信息查找并调用“方法表”中定义的方法. 进而根据方法返回参数进行推理. 推理机针对场景内容引用知识库中的知识进行推理. 直到知识扫描完毕.

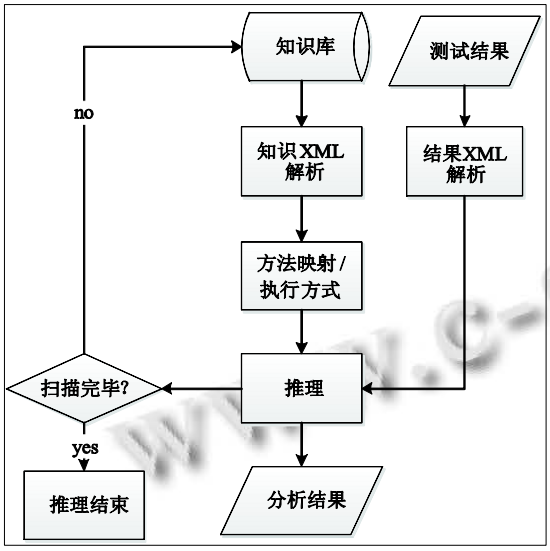

图 5 推理机结构

6 专家系统要面对的问题

在专家系统的建立及应用过程中主要面对的问题有以下几点:

① 知识储备

知识是专家系统运行的基础, 知识储备量决定了专家系统的分析能力. 如何有效的扩充分析专家系统的知识量, 以及确保专家系统的知识质量是专家系统在日常应用中要面对的重要问题.

② 分析方法的扩充

分析方法是值推理机在工作工程中应用到的推理动作. 每项知识内容都需要应用相应的推理方法. 对于新知识可能要应用到新的推理方法. 所以专家系统需要开发人员进行维护, 以确保能够适应提交的知识内容.

7 结论

本文提出了针对性能测试数据分析过程的专家系统. 并分别从数据表示方法、知识表示方法、推理机结构三方面简单的讨论了一种性能测试结果分析专家系统的构建和实现方式. 最后依据构建方式进一步讨论了专家系统的特性以及面临的问题.

参考文献

1 李怡,周国祥.基于 Load Runner 的一种性能测试流程方案研究与设计.计算机应用研究,2009,26(11):4143–4145.

2 IEEE Standards Board. IEEE Standard for Software Unit Testing: An American National Standard, ANSI/IEEE Std 1008-1987. IEEE Standards: Software Engineering,Volume Two: Process Standards, 1999 Edition.

3 Leszak M, Perry DE, Stoll D. A case study in root cause defect analysis. Proc. of the 22nd International Conference on Software Engineering (ICSE’00). 2000: 428–437.

(上接第 22 页)

报,2011,51(10).

6 梁斌.走进搜索引擎.北京:电子工业出版社,2007.

7 McCallum A, Nigam K, Lyle H. Ungar: efficient clustering of high-dimensional data sets with application to reference matching. Proc. of the 6th ACM SIGKDD. 2000. 169–178.

8 McQueen J. Some methods for classification and analysis of multivariate observations. Proc. of the 5th Berkeley Symp. on Math. Stat. and Prob. 1967,1:281–296.

9 周涛,柏文洁,汪秉宏,等.复杂网络研究概述.物理,2005,34(1):31–36.

10 陈旭.基于用户行为及关系的社交网络节点影响力评价—以微博研究为例.北京:北京邮电大学,2011.

11 620734263.hadoop 和 hbase 分布式配置及整合 eclipse 开发.[2011-07-20].http://wenku.baidu.com/view/8712a661caaedd3383c4d392.html

12 White T. Hadoop: The Definitive Guide: O'Reilly Media, 2009.

13 George L. HBase: The Definitive Guide: O'Reilly Media, 2011.

14 项斌.网络舆情检测系统设计与实现.成都:电子科技大学,2010.

15 陈旭.基于社会网络的 WEB 舆情系统的研究与实现.成都:电子科技大学,2010.

16 何忠育.分布式社会网络分析支撑系统研究与应用.广州:广东工业大学,2011.