第 33 卷第 2 期　　河 北 科 技 大 学 学 报　　Vol. 33, No. 2
2012 年 4 月　　Journal of Hebei University of Science and Technology　　Apr. 2012

文章编号:1008-1542(2012)02-0161-05

基于云计算的商业情报采集系统

许云峰[1],张　妍[1],赵铁军[2]

(1. 河北科技大学信息科学与工程学院,河北石家庄　050018;2. 河北省通信建设有限公司,河北石家庄　050021)

摘　要:商业情报采集系统不同于传统的搜索引擎系统,情报具有时效性、针对性等特点,传统搜索引擎中的数据分类和聚类技术不能完全满足商业情报采集过程中对时效性和针对性的特殊需求。提出一种商业情报采集解决方案,在云计算环境中采用贝叶斯分类算法和多种网页去重、提取等算法,实现对互联网数据的实时性抓取、分析、分类、聚类,形成对用户全方位立体化的情报本体,抓取的海量数据采用分布式文件系统存储,采集的情报用基于云的数据库 CouchDB 存储。

关键词:情报采集;搜索引擎;分类;聚类;云计算

中图分类号:TP391.1　　**文献标志码**:A

Cloud-based business intelligence gathering system

XU Yun-feng[1], ZHANG Yan[1], ZHAO Tie-jun[2]

(1. College of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China; 2. Hebei Communication Construction Company Limited, Shijiazhuang Hebei 050021, China)

Abstract: The business intelligence gathering system is different from the traditional search engine system. The data classification and clustering techniques of the traditional search engine can not fully meet the special needs of timeliness and pertinence in the business intelligence gathering process. This paper presents a solution to business intelligence gathering, by using Bayesian classification algorithm and deleting duplicated web pages algorithms in the cloud computing environment to achieve internet data's real-time capturing, analysis, classification and clustering, and form the omnibearing and three-dimensional intelligence noumenon of users. The amount of data captured is stored in a distributed file system. The gathered information is stored in the cloud database CouchDB.

Key words: intelligence gathering; search engine; classification; clustering; cloud computing

互联网数据浩如烟海,信息瞬息万变,如何在其中找到有价值的商业情报,不仅需要用户具备良好的情报意识,更重要的是拥有得力的搜索工具,这样可以使情报采集工作事半功倍。由于商业情报具有针对性、时效性等特点,传统搜索引擎技术中单纯的数据分类和聚类算法已不能满足商业情报搜索的需求[1]。笔者提出一种商业情报采集解决方案,可满足情报采集系统中对信息的时效性和针对性的需求。

1　系统设计

1.1　系统部署

系统部署见图 1。实时情报采集系统服务器集群部分由 8 台服务器构成,每个节点上都运行 FreeBSD

收稿日期:2011-11-04;责任编辑:陈书欣
基金项目:河北省科技支撑计划资助项目(10213588)
作者简介:许云峰(1980-),男,河北沧州人,讲师,硕士,主要从事网络安全、神经网络等方面的研究。

UNIX系统，并且配置了Hadoop环境。Hadoop是Apache软件的顶级项目，其包括很多子项目，如Hadoop Core，Hbase，Hive，pig，ZoomKeeper等[2]。其中Hadoop Core为利用普通PC硬件构建云计算环境提供基本服务，并且为开发云上的应用提供了基本API。Hadoop Core主要由HDFS和MapReduce模型构成。用户可以采用PC机、笔记本电脑和智能手机等多种手段获取情报信息，多个部门之间可以信息共享，情报互通，多地域、多终端之间一起采集情报。

1.2 技术架构

系统技术架构见图2。该系统由5个功能模块构成：分布式文件系统和MapReduce框架、Web页面采集、信息处理、文件索引、情报模式库构建。其中分布式文件系统和MapReduce框架模块主要是架构在Hadoop平台上。Hadoop平台包括HDFS和MapReduce两大部分[3]。这些基本架构保证了该系统具有高容错性及对数据读写的高吞吐率，能自动处理失败节点。Web页面采集模块采用多主机、多线程并行下载，将网页保存到HDFS信息处理模块并将网页预处理，提取核心文本进行文档分类，识别有价值情报信息，然后情报入库。根据输入的情报样本库信息，情报模式库模块提取特征模式存入模式库，供信息处理模块调用。文本索引模块对情报库CouchDB进行索引，然后供用户查询。该系统在云计算环境中实现了贝叶斯分类算法和多种网页去重、提取等算法，实现对互联网数据的实时性抓取、分析、分类、聚类，形成对用户全方位立体化的情报本体，同时抓取的海量数据采用分布式文件系统存储，采集的情报用基于云的数据库CouchDB存储。因采集的情报信息并不是结构化数据，且互联网数据良莠不齐，所以信息存储在传统的关系型数据库里，就会在存取过程中有很多麻烦。而采用CouchDB数据库存储这些情报信息就非常方便，而且增删改查更加便捷。另外采用CouchDB数据库可以让更多的客户端和服务器采用http访问情报库，从而提高管理和维护的效率。

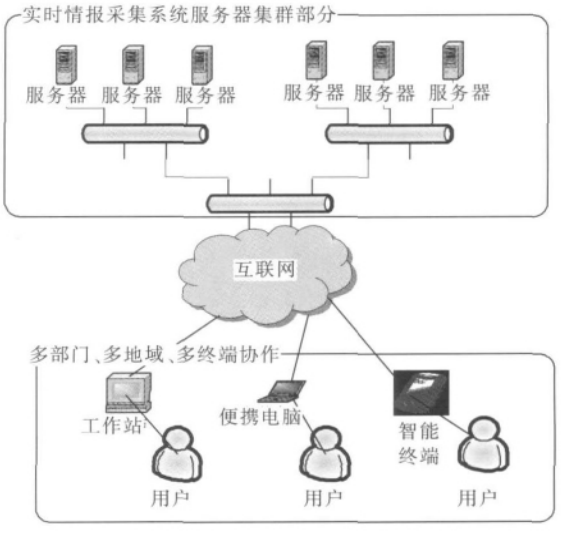

图1 基于云计算的情报采集系统部署

Fig. 1 Cloud-based intelligence gathering system deployment diagram

2 关键技术

本系统关键技术是让传统搜索引擎中常用的网页抓取、分析、分类、聚类等技术手段，在云计算环境中得以实现，提高运算速度并且满足了情报采集对时效性、针对性的需求，甚至可以达到情报信息采集的实时性。而对中文分词、分类、聚类的MapReduce化（即云环境下的算法实现），更是本技术的关键。

2.1 中文分词的MapReduce化

中文分词的MapReduce化的关键是中文分词组件在云计算环境里的分发。笔者选用IKAnalyzer这个开源的JAVA中文分词工具包，将其通过JobClient分发到Hadoop的各个节点中去。将IKAnalyzer的jar包解压缩，然后把它和源程序的类文件打包到一个jar包中。这种方法可以用Eclipse的Export功能轻松实现。

2.2 分类算法的MapReduce化

系统主要采用BAYES分类算法。分类器的输入是经过处理后的核心文本，为了唯一确定这些文本，文本的名字是网页的URL。

MapReduce化的流程是：1）核心文本的预处理，处理成由多行“URL\t 核心文本”构成的一个大文件；2）在Map中对核心文本进行分类，最后获得URL和分类的键值对。Map（URL，核心文本）→（URL，分类名）；3）在Reduce中对URL和分类名键值对进行统计。

2.3 聚类算法的MapReduce化

本文中数据聚类采用k-means算法。该算法的MapReduce化分为4个阶段：核心文本预处理；Map阶

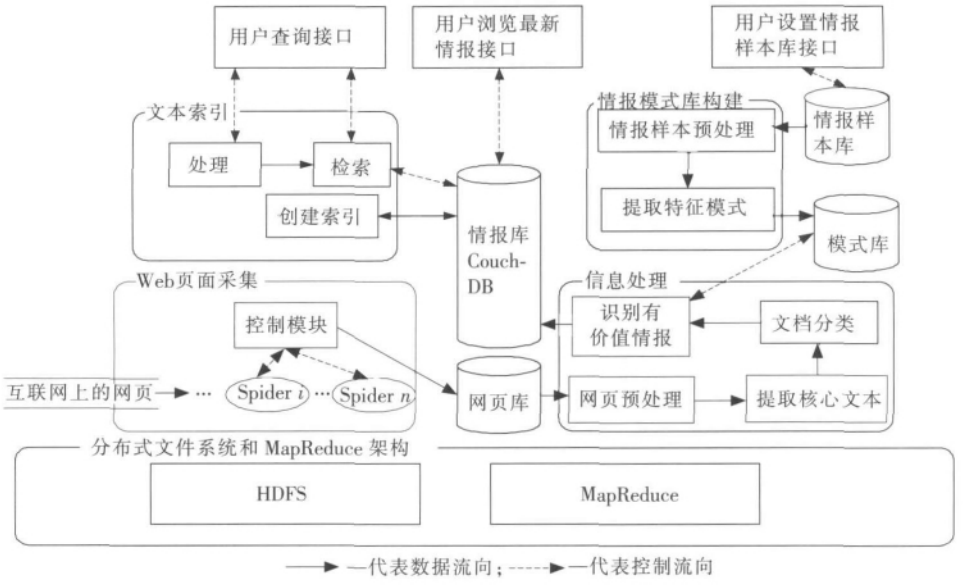

图 2　系统技术架构图

Fig. 2　System technical architecture diagram

段；Combine 阶段；Reduce 阶段[4]。并且文中还会用到 TF/IDF 权重，用余弦夹角计算文本相似度，用方差计算 2 个数据间欧式距离等相关算法[5]，鉴于篇幅有限，在这里不再赘述。

1）核心文本预处理：首先根据核心文本生成一个多行的“URL\t term1 TF/IDF；term2 TF/IDF；term3 TF/IDF；…”构成的大文件。

2）Map 阶段：构建 1 个全局变量 Centers，长度为用户选择的分类数，并把中心点赋给 Centers，中心点为随机选中的任何一行文本。Map 程序调用用户自定义的 SequenceFileInputFormat 读取 value 值。计算每一行文本与每一个聚类中心的距离，并将最小距离的聚类 ID 保存下来。构造〈key，value〉的形式传递给 Combine 阶段的处理，其中 key 是聚类中心点离每行文本的 value 值最近的聚类 ID。

3）Combine 阶段：首先初始化一个向量类型的 NewCenter 来储存新的中心点。将局部的具有相同聚类 ID 的文本数据进行距离比较，选择新的中心点，然后存储到 NewCenter 中。构造〈key，value〉的形式传递给 Reduce 阶段的处理，其中 key 为聚类 ID。

4）Reduce 阶段：首先初始化一个向量类型的 NewCenter_all 来储存新的中心点。将所有节点的具有相同聚类 ID 的文本数据进行距离比较，选择新的中心点，然后存储到 NewCenter_all 中。构造〈key，value〉，其中 key 为聚类 ID，value 为新的中心点。

5）判断每一行文本数据是否离唯一的聚类中心点距离最近，即判断是否收敛，如果是则程序退出，聚类完成，否则返回第 2 步，继续执行，直到收敛为止。

整个流程可以用图 3 表示。

2.4　情报库的云数据库存取

由于采集的情报信息并不是结构化数据，字段长度不一，所以数据存储在传统的关系型数据库里会使存取过程中有很多麻烦，另外 8 个节点并行采集的数据在存储时对数据库的并发连接可以达到数百个，这样传统的关系型数据库已不能满足存取需求[6]。本系统采用云架构的数据库 CouchDB，该数据库采用 Erlang 并行运算语言开发，特点就是支持并发数据连接，同时数据节点可根据需求轻松扩展，并且 CouchDB 中的数据记录可以是由任意个字段构成，因此采用 CouchDB 数据库存储情报信息非常方便。另外采用 CouchDB 数据库可以让更多的客户端和服务器采用 http 进行访问，从而提高管理和维护的效率。

3　系统测试结果分析

3.1　云环境下情报采集速度测试

测试环境：Dell R410 服务器 2 台，每台分别部署 Vmware vSphere Hypervisor。在每台服务器上，分别

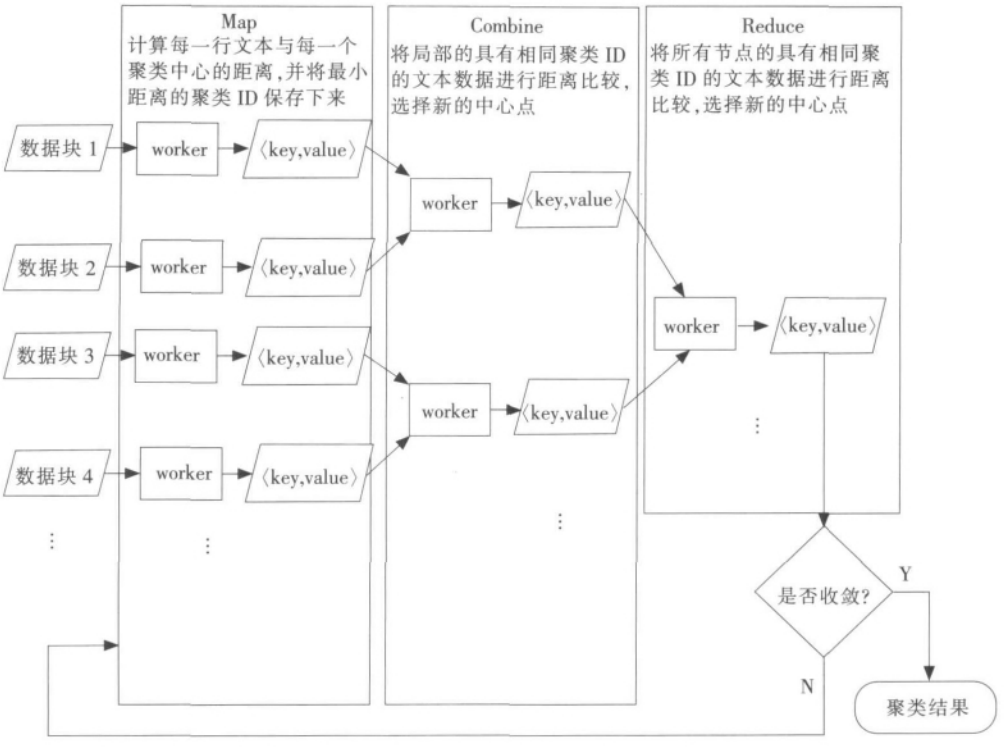

图3 聚类算法的MapReduce化

Fig. 3 Clustering algorithm based MapReduce

安装4个FreeBSD操作系统。在每个FreeBSD系统上面，部署安装Hadoop0.2。Dell R410的硬件配置：CPU Intel Xeon E5504，2 GHz主频，4 GB内存。测试环境架构见图4。

笔者对18 103个中文文本进行情报采集，首先进行中文分词，然后进行分类，再将分类识别后的有价值情报信息进行文本聚类，以符合用户的需求。测试中分别用5，6，7，8个节点的云计算环境，整个流程的运行时间依次为798 s，674 s，594 s，618 s。可见随着节点的增加，运行相同应用程序的时间线性减少。但是到8个节点的时候，运行时间有所加长。这是因为文中采用的是虚拟节点，因此网络开销就会加大，造成在8个节点时运行时间有所加长。如果采用物理节点，这个现象出现的可能性会变小。

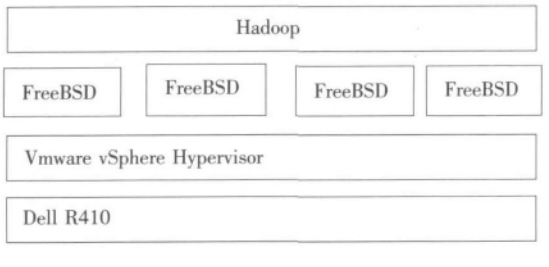

图4 测试环境架构图

Fig. 4 System technical architecture diagram

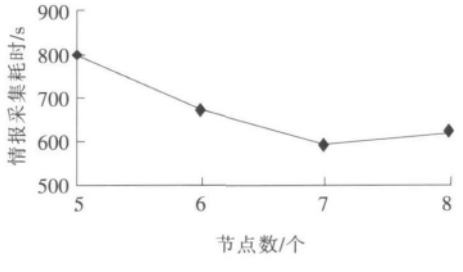

图5 节点和运行时间的关系图

Fig. 5 Node and run-time diagram

3.2 云环境下情报存取速度测试

云存储实验中，采用2台相同配置的服务器Dell R410，分别安装Mysql和CouchDB，对这2个库进行插入实验，将分类后有价值的情报100万条分别插入Mysql和CouchDB。Mysql耗时为79.965 s，占用2.3 GB左右的存储空间，CouchDB耗时为58.362 s，占用1.8 GB左右的存储空间。可见CouchDB比Mysql的插入效率要高，占用存储空间比Mysql要少。如果存入的100万条数据是字段统一的，Mysql要比CouchDB耗时要少，但是在插入Mysql前要根据数据库中表的字段对数据进行格式化，这个过程相当耗费系统时

间，所以导致 Mysql 最终要比 CouchDB 耗时多。

4　结　语

采用云计算技术，随着节点增加，情报信息的采集速度应该是线性提高的，当节点达到一定的数量时，可以满足情报信息对时效性、针对性的需求，甚至可以达到情报信息采集的实时性。

采用基于云计算的数据库 CouchDB 可以解决情报信息的非结构化问题，同时 CouchDB 数据库的节点可以根据需求不断扩展，满足用户对信息存取速度的需求。

参考文献：

[1] 张立岩，吕　玲. 基于最大熵算法的全文检索研究[J]. 河北科技大学学报(Journal of Hebei Universty of Science and Technology)，2009，30(2)：112-115.

[2] 林清滢. 基于 Hadoop 的云计算模型 [J]. 现代计算机(Modern Computer)，2010(7)：114-116.

[3] 周轶男，王　宇. Hadoop 文件系统性能分析[J]. 电子技术(Electronic Technology)，2011(5)：15-16.

[4] 江小平，李成华. *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)(Journal of Huazhong University of Science and Technology(Natural Science Edition))，2011，39(S1)：120-124.

[5] 赵卫中，马慧芳. 基于云计算平台 Hadoop 的并行 *k*-means 聚类算法设计研究[J]. 计算机科学(Computer Science)，2011，38(10)：166-168.

[6] 郝　伟，杨国霞. 专业搜索引擎搜索结果融合算法研究[J]. 河北科技大学学报(Journal of Hebei Universty of Science and Technology)，2011，32(4)：355-358.

(上接第 134 页)

对试验过程和数据进行记录，结果见表 3。

表 3　灭火试验结果

Tab. 3　Outfire test results

项　目	结　果
灭 A 类火灾效能	成功，灭火面积 0.6 m^2
灭 B 类火灾效能	成功，灭火面积 0.5 m^2
喷放时间/s	＜0.1
灭火时间/s	＜1
结　论	具有良好的灭 A，B 类火灾效果

5　结　语

根据设计方案，初步制作了警用战术灭火弹样弹，并进行了相应的试验，证明该弹在战场条件下的初期火灾时具有良好的扑灭效果；同时，喷出的超细干粉灭火剂可在一定区域内形成烟雾屏障，能遮挡歹徒视线，利于官兵快速制敌，满足部队处置突发事件的战术要求。

参考文献：

[1] 杜力强，柴　涛. 超细干粉灭火技术探讨[J]. 机械管理开发(Mechanical Management and Development)，2008，23(6)：93-95.

[2] 吴颐伦. 干粉灭火剂配方一般原理[J]. 消防技术与产品信息(Fire Technique and Products Information)，2000，19(6)：19-25.

[3] 杨　杰. 灭火气溶胶发生剂灭火机理及配方设计[J]. 火炸药学报(Chinese Journal of Explosives & Propellants)，2003，26(4)：84-86.

[4] 刘玉梅，张文超，潘仁明. 气溶胶灭火技术的研究现状和发展趋势[J]. 消防技术与产品信息(Fire Technique and Products Information)，2003，22(4)：69-70.

[5] HEINONEN E W，TAPSCOTT R，KIBERT C J，et al. Aerosol technology overview and bibliography[R]. [S. l.]：Marine and Fisheries Engineering Research Inst Inc Woods Hole Ma，1995.

[6] 蔡瑞娇. 火工品设计原理[M]. 北京：北京理工大学出版社，1997.

[7] 何以申. 细水雾的灭火机理和功能认证[J]. 消防技术与产品信息(Fire Technique and Products Information)，2009，28(2)：36-37.

[8] 丁玉兰. 人机工程学[M]. 北京：北京理工大学出版社，2005.

[9] 李习民，谢玖伟，秦玉旺，等. 非贮压干粉灭火装置的改进[J]. 消防科学与技术(Fire Science and Technology)，2010，29(9)：790-793.

[10] 武警工程学院训练部. 枪械理论[R]. 西安：武警工程学院，2010.