Computer Science and Application 计算机科学与应用, 2016, 6(7), 443-450
Published Online July 2016 in Hans. http://www.hanspub.org/journal/csa
http://dx.doi.org/10.12677/csa.2016.67054

Hans汉斯

Research on Naive Bayesian Spam SMS Filtering Based on MapReduce

Caidi Zhao, Youchan Zhu, Jiahui Fu

North China Electric Power University, Baoding Hebei
Email: zcd_cd@163.com

Received: Jul. 8th, 2016; accepted: Jul. 26th, 2016; published: Jul. 29th, 2016

Abstract

The massive text mining filter requires a lot of storage space and stronger computing ability, so a spam message filtering method of MapReduce-based Bayesian is proposed. Based on the improved Naive Bayesian spam SMS classification algorithm, taking the advantage of MapReduce model parallelization on massive data processing is used to train and test SMS text. Results show that using compute cluster to achieve massive spam filtering can improve the efficiency of recalling and precision, and with the expansion of cluster size spam SMS filtering efficiency improve faster.

Keywords

Spam SMS, SMS Filter, Naive Bayesian, MapReduce

基于MapReduce的朴素贝叶斯垃圾短信过滤研究

赵彩迪，朱有产，符佳慧

华北电力大学，河北 保定
Email: zcd_cd@163.com

收稿日期：2016年7月8日；录用日期：2016年7月26日；发布日期：2016年7月29日

文章引用: 赵彩迪, 朱有产, 符佳慧. 基于 MapReduce 的朴素贝叶斯垃圾短信过滤研究[J]. 计算机科学与应用, 2016, 6(7): 443-450. http://dx.doi.org/10.12677/csa.2016.67054

摘　要

针对海量短信文本的挖掘过滤需要很大的存储空间以及更强的计算能力，提出一种基于MapReduce的朴素贝叶斯的垃圾短信过滤方法；基于改进的朴素贝叶斯垃圾短信分类算法，利用MapReduce模型并行化对海量数据处理的优势进行短信文本的训练和测试。实验表明：利用计算集群实现海量垃圾短信过滤在召回率、查准率方面有所提高，垃圾短信过滤效率随着集群规模的扩增而提升较快。

关键词

垃圾短信，短信过滤，朴素贝叶斯，MapReduce

1. 引言

由于手机普及率的提高和短信通信费的低廉，垃圾短信已经严重侵扰到了手机用户的正常生活，诈骗短信更是使不少用户蒙受损失。《2015上半年中国移动互联网安全报告》显示，全国垃圾短信数量高达199亿条。所以为广大用户建立起来一个可靠、准确、高效、智能的短信过滤平台，对手机短信实施有效的管制，具有重要的意义和价值。

当前垃圾短信过滤技术主要分为基于关键词和基于短信内容的过滤。前者要求只要短信中包括的敏感词汇超过一定数目就被认定为垃圾短信，该方法简单但是误判率比较高；一种是基于短信内容的过滤，通过机器学习的方法把短信自动分为正常短信和垃圾短信。通过对基于机器学习的短信文本自动分类[1]常用算法朴素贝叶斯[2]、SVM[3]、KNN[4]、人工神经网络算法等有缺点进行分析，选取简单高效的朴素贝叶斯算法。针对海量短信数据，本文提出基于MapReduce的朴素贝叶斯垃圾短信过滤方案，将垃圾短信过滤过程移植到云计算平台，充分利用云计算平台对海量数据的高效存储和处理能力，通过MapReduce模型对垃圾短信过滤过程进行并行化处理，实现垃圾短信高效、智能的过滤。

2. 关键技术

2.1. MapReduce 模型

MapReduce模型[5]是一种能在大型计算机集群上并行处理海量数据的框架模型，是Hadoop的三大核心技术之一，由Google公司首先提出。MapReduce模型作为一种简化的并行计算模型，编程模型借鉴了函数式语言中map和reduce函数的启发，将数据处理过程抽象为Map阶段和Reduce阶段：

(1) 在Map 阶段，MapReduce模型接受键值对<key,value>作为数据输入，经过映射，聚合所有具有相同的 key 值的中间结果的value值，产生一组键值对<key1, value1>形式的中间结果。

(2) 在Reduce 阶段，以Map的中间结果作为Reduce函数的输入，把具有相同Key值的中间结果进行汇总处理，输出最终结果<key2, value2>。

整个 Map/Reduce 框架下处理海量数据的流程图如图 1 所示。

2.2. 短信特征提取

除了分类算法，特征提取很大程度上影响文本分类效果。特征提取[6]主要通过某个特征评估函数映射(或变换)的方法，选择代表意义较强、分类性较好的特征项进行文本表示，组合成文本向量。常用的特征提取方法有TF-IDF，信息增益(IG)，互信息(MI)，卡方统计量方法等。经过分析，本文选择TF-IDF

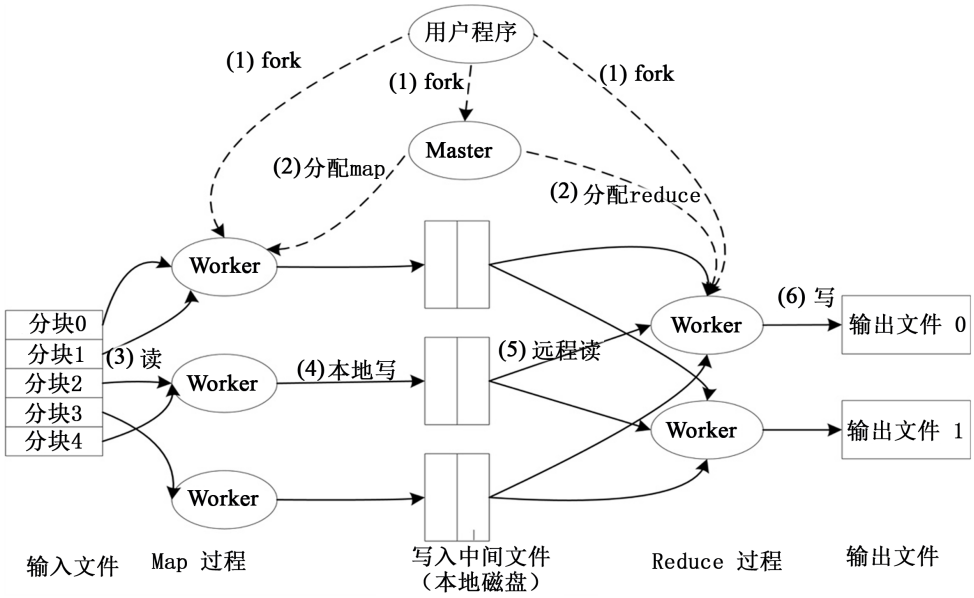


Figure 1. The flow chart of processing massive data with MapReduce
图 1. MapRedeuce 处理海量数据流程图

作为特征提取的标准[7]，根据TF-IDF算法度量文本中每个词项的权重，挑选具有较高权重的词项作为文本的特征次。词语权重计算是利用tf和idf的乘积来实现的：Weight ＝ tf * idf。

基于TF-IDF作为特征提取的标准[8]，创造性提出一种基于词项多种因子的特征次选择算法，将词项的TF-IDF、词性与词长因子构造综合评估函数。如此所抽取出的特征词较不加任何因素的TF-IDF方法更能准确地表征文本内容，同时也克服TF-IDF等算法无法解决的高频但无实际含义词项的误判问题，提高了文本特征词选取的准确率。基于词项TF-DF、词性以及词长等因子计算表征文本D中每一个词项term，重要性的总权重分值，如式(1)所示。

$$\mathrm{Score}(\mathrm{term_i}) = \alpha \times \mathrm{Weight_{tf-idf}}(\mathrm{term_i}) + \beta \times \mathrm{W_{Pos}}(\mathrm{term_i}) + \gamma \times \mathrm{W_{Length}}(\mathrm{term_i}) \quad (1)$$

其中，$\mathrm{term_i}$ 表示当前文本D中的第i个词项，$\mathrm{Weight_{tf-idf}}(\mathrm{term_i})$ 表示词项对应的TF-IDF权重分值，$\mathrm{W_{Pos}}(\mathrm{term_i})$ 表示词性权重分值，$\mathrm{W_{Length}}(\mathrm{term_i})$ 表示词长权重分值，而 α, β, γ 则表示词项$\mathrm{term_i}$的不同因子在决定其在文本中重要性的比例系数。

3. 朴素贝叶斯短信文本分类

由于分类算法中SVM算法复杂程度过高，不适用于大规模的文本，本文采用经典文本分类算法朴素贝叶斯算法[9]。基于改进的朴素贝叶斯短信文本分类过程：

(1) 将短信文本表示为特征向量 $T = \{t_1, t_2, \cdots, t_k, \cdots, t_n\}$，其中，$t_k$表示第$k$个特征项，表示特征向量空间的一个向量

(2) 定义短信类别 $D = (D_1, D_2, \cdots, D_n)$，本文定义两种类别包括正常短信和垃圾短信。某个未知类型的短信的特征向量T，用贝叶斯公式计算T属于类 D_i 的概率。

$$P(D_i \mid T) = \frac{P(T \mid D_i) P(D_i)}{P(T)} \quad (2)$$

在上面公式中 $P(D_i \mid T)$ 表示在T条件下 D_i 发生的概率,即T属于 D_i 类的概率。$P(D_i)$ 表示 D_i 类的先验

概率，即 $P(T|D_i)$ 表示T在 D_i 类出现的概率，$P(T)$表示特征的先验概率。

(3) 假设短信文本之间不存在依赖关系，即样本的特征向量之间是独立的，计算 $P(T|D_i)$，由式子(3)计算出联合特征向量下的概率

$$P(T|D_i)P(D_i)=P(D_i)\prod_j^m P(t_i|D_i) \tag{3}$$

最后计算出 $P(T|D_i)P(D_i)$ 的最大项作为最终分类。

(4) 当有一个未知类别的短信样本T，贝叶斯分类法预测T属于具有最高后验概率的类 D_i，其成立的条件为：当且仅当 $P(D_i|T)>P(D_j|T)$，且$1\le j\le 2$，$j\ne i$。

(5) 当利用朴素贝叶斯分类器对短信进行分类时，易产生分类错误,通常用户可以接受将垃圾短信判定为合法短信,但是不能接受将合法短信被误判为垃圾短信。设定一个阈值R对朴素贝叶斯方法进行改进,令 $K=P(D_{Spam}|T)/P(D_{Ham}|T)$，当$K$大于$R$时，分类器将样本$T$识别为垃圾短信。

4. 基于 MapReduce 的朴素贝叶斯垃圾短信过滤设计

4.1. 方案总体设计图

垃圾短信过滤分类器的设计由训练过滤部分和过滤部分组成[10]，朴素贝叶斯垃圾短信文本分类流程图如下图 2 所示。

根据算法流程图，将训练模块和过滤模块共有的短信文本预处理部分抽离出来，独立成预处理模块，提高代码的重用效率，将预处理部分抽离出来。基于预处理部分耗费大量的时间和计算资源，本文对预处理部分进行 Mapreduce 并行化改进，预处理部分 MapReduce 模型如表 1。

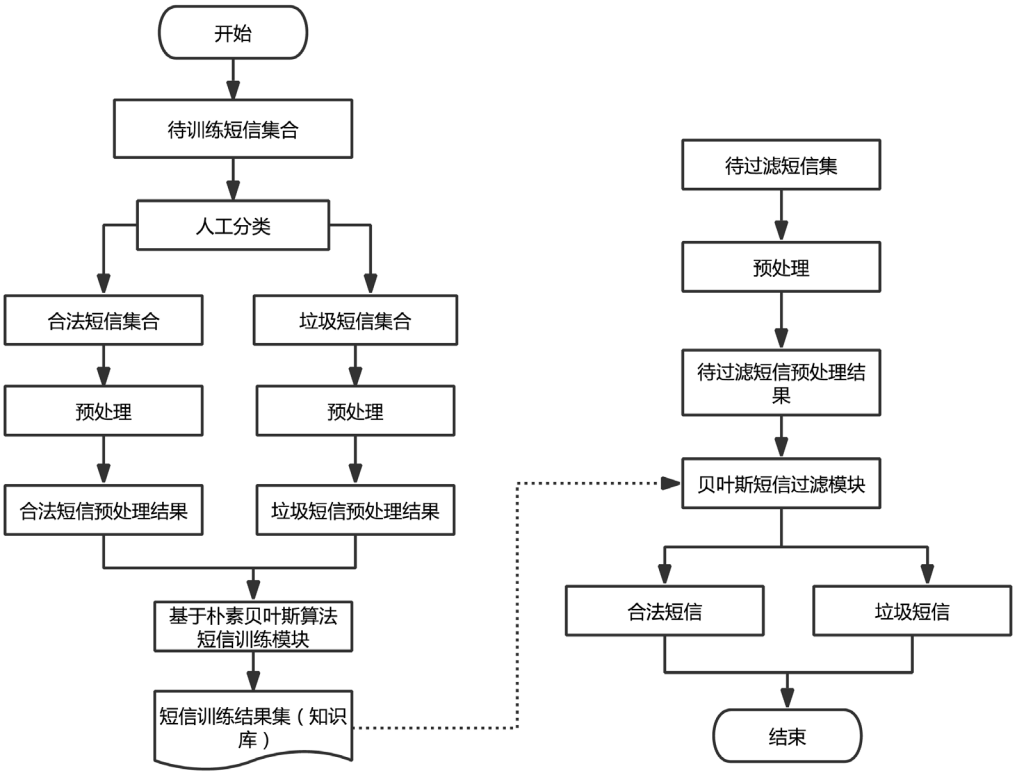


Figure 2. The flowchart of Naive Bayesian spam SMS text classification
图 2. 朴素贝叶斯垃圾短信文本分类流程图

Table 1. The MapReduce model of preprocessing module
表1. 预处理模块MapReduce模型

MapReduce	功能	对需要预处理短信文本内容进行分词等处理
	输入	需要预处理短信文本内容
	输出	预处理后的邮件特征向量的词汇总数及每个词汇向量的频率

4.2. 分类器构建

目前挖掘分类技术[11]只能处理结构化的数据，将短信文本数据进行预处理，从非结构化数据变成了结构化数据进行挖掘分类。在云计算的技术环境下主要分三个步骤来完成分类器的训练工作：

步骤一：文本的分布式预处理

将待处理短信文本存储在分布式文件系统HDFS的不同DataNode中，本文选择适合文本分类的方法Sequence File，将每个短信小文本按类别归并为大文件以键值对的形式存储，每个文本占据一行，添加一个全局标号fileId。之后采用中科院的ICTCLAS分词系统，对文本进行分词，去除停用词、词频统计，最中生成排序引文件。

步骤二：文本的分布式特征选择

利用MapReduce对特征选择过程进行分布式并行设计，本文采用基于词项多种因子的特征次选择算法，将词项的TF-IDF、词性与词长因子构造综合评估函数，以此来选择代表意义较强、分类性较好的特征项进行文本表示，组合成新的文本向量。如此所抽取出的特征词较不加任何因素的TF-IDF方法更能准确地表征文本内容，同时也克服TF-IDF等算法无法解决的高频但无实际含义词项的误判问题，提高了文本特征词选取的准确率。

步骤三：贝叶斯短信文本过滤模型的分布式并行训练

通过对朴素贝叶斯算法和并行策略的分析，在Hadoop平台上设计了基于朴素贝叶斯算法的短信训练模型，构件朴素贝叶斯分类器。

4.3. 分类器测试

通过分类器对短信文本测试过程采用两轮 MapReduce 的方法，流程图如图 3 所示。

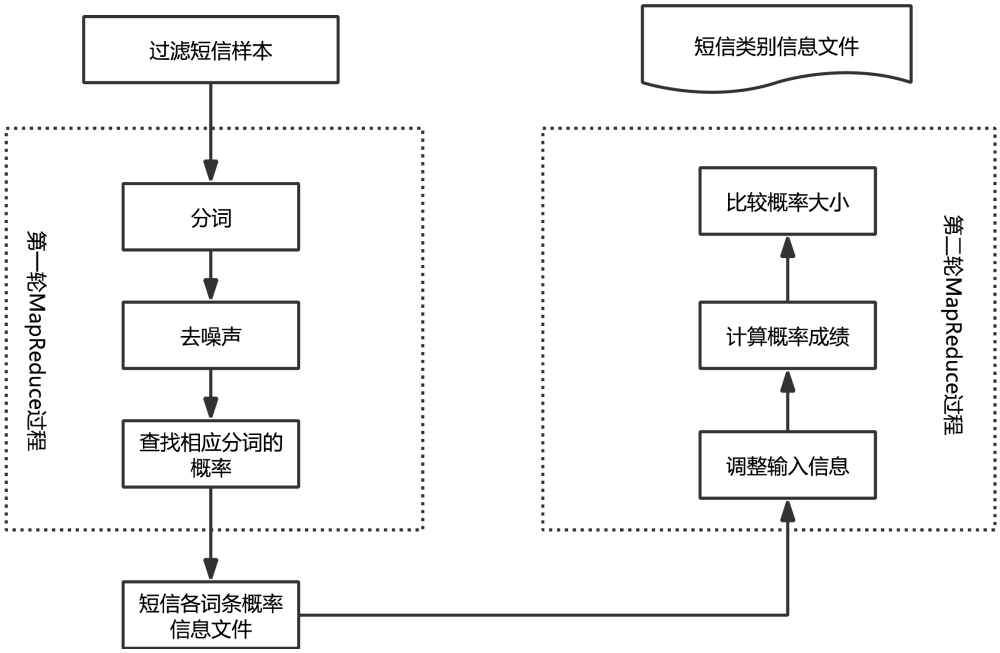
第一轮MapReduce对短信进行分词和去除噪声。每个Map函数接收一个短信数据块，输入键值对：<key (偏移量 + 主题)，value (短信数据块)>，经过分词处理，输出中间结果<key (类别丨词条)，value (词频)>。每个Reduce函数接受Map的输出结果，合并具有相同Key的value 值，得到各词条的数量统计以及计算得各词条的概率，输出结果为< key(类别丨词条)，value(“合法概率|垃圾概率”)>键值对，从而产生相应分词的词条计数结果文件，供第二轮MapReduce使用。

第二轮MapReduce计算概率比较大小。Map阶段是对输入键值对< key(类别丨词条)，value(词频)>的格式转换，通过拆分输入信息Key得到短信的标识，将一条短信的各个分词集中一块，并以此做输出的Key。Reduce阶段运算相应的结果，输出格式为<Key(短信的Key)，Value(T/F)>，如果Value 值为T则表示是合法短信，否则为垃圾短信。

5. 实验结果和分析

5.1. 实验环境搭建

在实验室局域网环境下搭建了包含 5 个节点的云集群，各主机的机器名及分配的 IP 地址如表 2 所示。



Figue 3. The testing process of SMS text classifier
图 3. 分类器短信文本测试流程

Table 2. Cloud cluster host machine name and IP address allocation table
表2. 云集群主机机器名及IP地址分配表

功能	机器名	IP 地址	Tag 标记
Namenode (jobtracker)	hadoop.master	202.206.219.60	A
Datanode (tasktracker)	hadoop.slave61	202.206.219.61	B
Datanode (tasktracker)	hadoop.slave62	202.206.219.62	B
Datanode (tasktracker)	hadoop.slave63	202.206.219.63	B
Datanode (tasktracker)	hadoop.slave64	202.206.219.64	B

本课题实验数据采用中国移动通信研究院所提供的人工标注的数据,其中训练集包含样本短信30000条,测试集包含短信30,000条。实验在两种环境下进行性能测试。实验A是在普通的PC单机上基于传统的朴素贝叶斯短信过滤算法进行短信学习和分类；实验B是在Hadoop集群(单节点)上基于MapReduce的朴素贝叶斯短信过滤算法进行短信学习和分类。

5.2. 实验评估标准

本课题采用查准率(Precision)，召回率(Recall)，加速比(Speed-up Ratio)作为对垃圾短信过滤器性能的评价指标。

查准率 P(Precision)：经分类器判断，被判为垃圾短信的所有短信中确实是垃圾短信的比例，反映过滤器辨识垃圾短信的能力。

召回率 R(Recall)：实际情况中所有属于垃圾短信的文本，经分类器判断，识别出的垃圾短信占实际垃圾短信总数的比例，反应过滤器的过滤能力。

加速比(Speed-up Ratio)：系统执行同一个任务，在系统改进前后所用时间的比例，反映系统执行效率的高低。

5.3. 实验结果分析

(1) 基于 MapReduce 的贝叶斯短信过虑算法与传统的朴素贝叶斯短信过滤算法的实验结果如下表 3 所示，在查准率和召回率上实验性能对比如下表 4 所示。

由表4的实验数据对比可以发现，基于MapReduce的改进的贝叶斯短信过滤算法与传统的朴素贝叶斯短信过滤算法在查准率和召回率上有所提高。

(2) 测试基于 MapReduce 的贝叶斯短信过滤算法与传统的朴素贝叶斯短信过滤算法的加速比数据如表 5，加速比性能对比如下图 4。

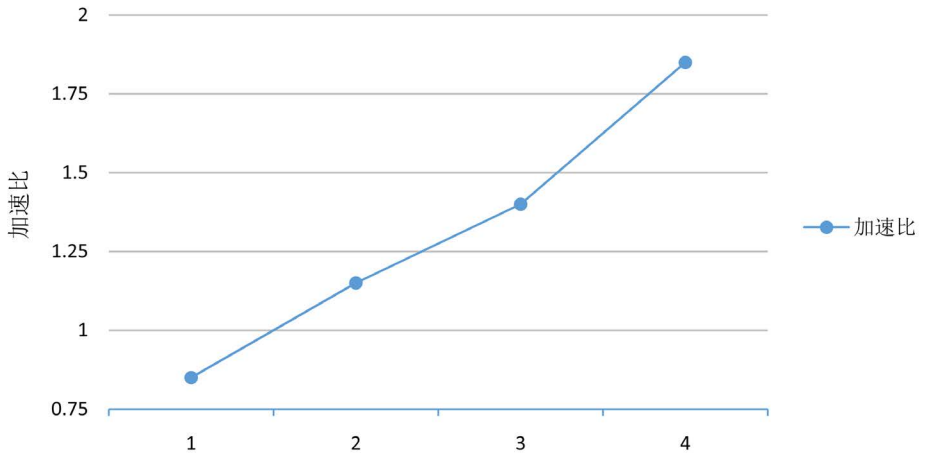


Figure 4. Linear speedup under different nodes
图 4. 不同节点下加速比线性图

Table 3. The results of the experiment
表3. 实验结果

实验组别	邮件类型	判为合法短信	判为垃圾短信
实验A	合法短信	7959	1781
	垃圾短信	2678	17582
实验B	合法短信	8471	1269
	垃圾短信	1589	18671

Table 4. Comparison of experimental performance
表4. 实验性能对比

实验组别	查准率/%	召回率/%
实验A	86.78	90.80
实验B	92.15	93.63

Table 5. Speedup of single machine and different nodes
表5. 单机以及不同节点时加速比

节点个数	单机	1	2	3	4
时间(s)	138.1	159	119.7	97.6	75.3
加速比		0.86	1.15	1.41	1.83

分析得知：

(1) 在一个节点的情况下，由于从节点本身的TaskTracker和DataNode等的运行开销，加速比小于1，反而不如单机程序效率高。

(2) 当大于一个节点时，随着节点树目增加，加速比随之增加，各阶段的加速比越来越大，说明各个阶段的MapReduce过程得到了分布式并行运行，作业的并发程度高，运行效率高。

6. 总结

经过软硬件平台的搭建和实验分析，本文提出的基于 MapReduce 朴素贝叶斯垃圾短信过滤方案，一方面对特征选择和朴素贝叶斯算法进行改进，提高了垃圾短信过滤的召回率和查准率；另一方面引入 MapReduce 模型，对分类过程进行并行化处理，提高短信文本学习和分类的过程。同时垃圾短信过滤的处理对其它的微博、留言等其它短文本的过滤具有借鉴意义。

参考文献 (References)

[1] 刘依璐. 基于机器学习的中文文本分类研究[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2009.

[2] Joachims, T. (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Feature. *Proceedings* of 10*th European Conference on Machine Learning*, New York.

[3] Cosatto, E., Bottou, L., Dourdanovic, I., *et al.* (2004) Parallel Support Vector Machines: The Cascade SVM. *Neural Information Processing Systems*, 2004.

[4] 李荣陆, 胡运发. 基于密度的 KNN 文本分类器训练样本裁剪方法[J]. 计算机研究与发展, 2004, 41(4): 539-545.

[5] Dean, J. and Ghemawat, S. (2008) MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, **51**, 107-113. http://dx.doi.org/10.1145/1327452.1327492

[6] 陈雨杰. 文本分类中特征选择算法研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2015.

[7] 施聪莺, 徐朝军, 杨晓江. TFIDF 算法研综述[J]. 计算机应用, 2009, 29(S1): 57-60.

[8] 张爱华, 靖红芳, 王斌, 等. 文本分类中特征权重因子的作用研究[J]. 中文信息学报, 2010, 24(3): 97-104.

[9] 江小平, 等. 云计算环境下朴素贝叶斯文本分类算法的实现[J]. 计算机应用, 2011, 31(9): 2551-2554.

[10] 朱杰. 云计算在基于贝叶斯分类的垃圾短信过滤中的研究与应用[D]: [硕士学位论文]. 成都: 电子科技大学, 2010.

[11] 何元. 基于云计算的海量数据挖掘分类算法研究[D]: [硕士学位论文]. 成都: 电子科技大学, 2011.