分布式环境下的文本聚类研究与实现

赵华茗
(中国科学院文献情报中心　北京 100190)

摘要: 【目的】通过开源工具, 构建一种分布式环境下的文本聚类与分类应用平台。【方法】以海量文本的词收敛性为基础, 通过词聚类指导文本聚类和分类。过程包括: 使用开源分词器等工具进行训练集的文本预处理, 结合 Mahout 数据挖掘平台对处理后的词集进行聚类分析, 最后通过相似度算法计算测试文本与词类簇的相似度并分类。【结果】分布式环境下的基于词聚类的文本聚类分类计算方法, 可有效解决海量文本的词聚类瓶颈问题。经测试, 当训练文本集增加到 100, 迭代收敛阈值为 0.01 时, 词聚类结果较理想。【局限】测试数据规模有限, 仅限于新闻数据, 基于其他领域的词聚类效果需要进一步测试、优化、调整。【结论】详细描述基于词聚类的文本聚类分类算法的开发环境构架和关键步骤, 有助于研究者对相关开源工具使用及分布式并行环境部署的深入理解。

关键词: 分布式环境　聚类　文本聚类　Hadoop　Mahout
分类号: TP393

1 引　言

随着互联网的发展, 数字信息呈几何级数增长。大多数数据是以文本非结构化的形式存储在计算机和网络上, 如何在海量同时蕴含着巨大潜在价值的未知信息中挖掘出有用的知识已经成为人们关注的热点问题之一。文本聚类是文本挖掘的一种有效方法, 它将文本集分为若干个子集, 使得类内的成员相似度尽可能大, 类间的成员相似度尽可能小。文本聚类可广泛应用于文本挖掘与信息检索的不同方面, 在大规模文本集的组织与浏览、文本集层次归类的自动生成等方面都具有重要的应用价值。

目前较常用的文本聚类算法有分层法、分割法、密度法、网格法、基于模型的方法等[1-8]。但这些算法多数都对文本特征矩阵的维度有较为苛刻的要求, 一旦维度增大, 则聚类效果会急剧下降。为了改进这种情况, Slonim 等[9]提出了使用单词聚类来指导文档聚类的方法。在该方法中表示文档的单词被提取出来, 然后计算它们属于一篇文档的概率, 再通过贝叶斯公式来计算两个单词属于同一个类的概率, 进而通过此概率对单词进行聚类; 然后再通过文档和单词类的关系对文档分类, 该方法较好地解决了降维问题, 但海量文本的词聚类过程仍是瓶颈。梁维铿[10]提出了基于 MapReduce[11]架构的文本聚类计算方法和思路, 通过设计 Map 函数和 Reduce 函数, 实现海量文本聚类计算的并行化, 但 MapReduce 编程过程较复杂, 也不利于算法的再次复用和维护修改。

针对上述问题, 本文利用 Mahout[12]大数据挖掘工具, 提出一种分布式环境下的基于词聚类的文本聚类分类计算方法, 有效解决海量文本的词聚类瓶颈问题, 并结合开源软件及开发框架, 详细阐述技术实现思路和关键点。

2 分布式环境下的文本聚类计算思路和关键技术

2.1 实现思路及步骤

分布式环境下, 实现文本聚类计算的总体思路是

通讯作者: 赵华茗, ORCID: 0000-0002-8829-9208, E-mail: zhaohm@mail.las.ac.cn。

以海量文本的词收敛性为基础，通过词聚类指导文本聚类和分类。系统按功能分为三个主要部分：训练集文本的预处理、词聚类和测试集文本分类计算，系统构架如图 1 所示：

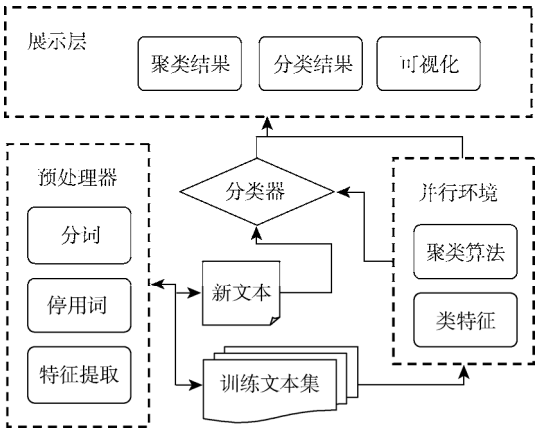

图 1　基于词类簇的文本聚类分类架构

具体实现过程如下：首先将存储在网络中的非结构化文本根据计算需要做清洗、分词过滤等文本预处理，去掉文本中的噪音数据，形成表示文档的词集；然后导入 Hadoop[13]分布式计算平台，结合 Mahout 数据挖掘工具，对清洗过的词集进行聚类，并通过阈值提取每个类中的前 N 个特征词表示该类，形成单词类簇；最后通过计算测试文本和单词类簇的相似度对测试文本分类。

2.2　关键技术

Mahout 是 Apache Software Foundation(ASF)旗下的一个开源项目，提供一些可扩展的机器学习领域经典算法的实现，旨在帮助开发人员更加方便快捷地创建智能应用程序[12]。Mahout 包含集群、分类、推荐过滤、频繁子项挖掘等算法实现，并实现了部分数据挖掘算法在 MapReduce 环境下的并行化。因此，通过使用 Apache Hadoop 库，Mahout 可以有效地扩展到分布式环境中。

3　分布式环境下的文本聚类计算的关键点

为了充分有效利用 Mahout 平台在大数据分析处理方面功能强大的优势，本文在海量非结构化文本聚类计算开发时，围绕 Mahout 平台，需仔细考虑文本训练集数据的预处理、聚类后特征词提取及测试文档的分类计算等过程和实现细节，也是整个文本聚类分类计算系统实现的关键点，文本聚类分类计算逻辑如图 2 所示：

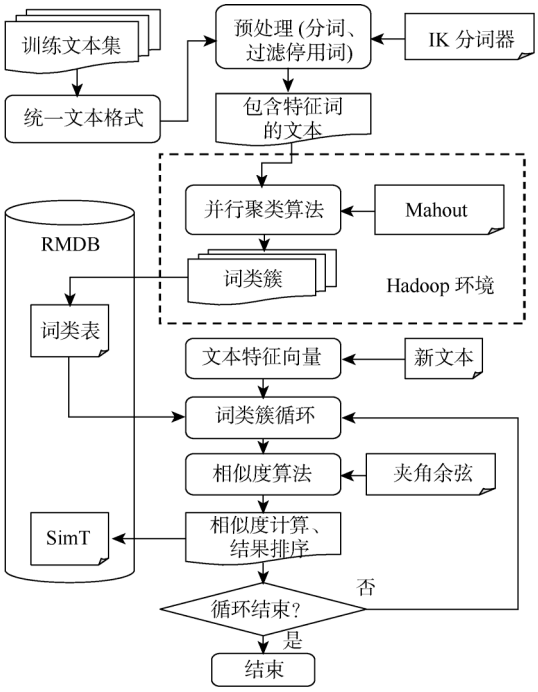

图 2　分布式的文本聚类分类逻辑

3.1　非结构化文本的数据预处理

为便于计算机处理，文本文档处理前必须要有一种有效的表示方式。在信息处理领域，向量空间模型(Vector Space Model, VSM)是应用较多且效果较好的表示方法之一[14]。在此模型中，每篇文档都表示成形如 D: <t1, w1; t2, w2; …; tn, wn>的向量，其中 ti 是词条项，wi 是 ti 在文档中的权值。因此，所有的 n 维词条向量组成了一个文档向量空间。

但使用特征词的向量空间来表示文档时，直接使用构成文档的词条作为向量空间的维度，会形成高维低密度矩阵，而且存在着大量对文档的描述和区分不相关或影响很小的词条维度，造成对文档语义描述的模糊。为了提高算法的效率和准确度，有必要对构成文本的词条进行特征词的提取和筛选，即对词条向量空间进行降维处理。

非结构化文本的预处理过程就是识别和提取非结构化文本特征词的过程，通过数据清洗和规范化处理，形成适合的中间格式，为数据导入 Hadoop 平台和后期数据挖掘做准备。考虑到 Mahout 数据挖掘工具本身提供的序列化及分词工具对中文的支持不

好，文本在数据预处理过程中使用 IK 中文分词器及相关的数据清洗字典，通过梳理数据字典和 IK 中文分词工具中的停用词字典，经过算法比对，提前过滤文本文档中的不相关词条或无意义词条，并完成分词。在数据预处理过程中，只对词条项 ti 做处理，权值 wi 交给数据挖掘过程处理。主要实现过程及伪代码如下：

```
If (数据源文件不为空并且格式正确) {
  For (iterator I =文件列表; i.hasNext){
        读文档;
        IK 分词器分词;
        通过数据字典过滤无意义及不相关词条;
        写文档;
  }
}结束
```

3.2 词聚类

预处理后的文本只包含表示该文本的特征词，系统将上传这些文本至 Hadoop 环境中交给 Mahout 工具进行聚类分析处理，并根据阈值，提取与类簇最相关的 N 个特征词，形成单词类簇。Mahout 工具提供了很多经典数据挖掘算法，主要的聚类算法包括：K-means、Canopy、Fuzzy-Kmeans、StreamingKmeans、Spectral Clustering 等 5 种[12]。其中，模糊 c-均值聚类(Fuzzy-Kmeans[15])较适合文本聚类，该算法是 Bezdek 于 1981 年提出的，是可重叠聚类算法，是 K 均值聚类的扩展，它的基本原理和 K 均值算法一样，只是它的聚类结果允许存在特征向量属于多个类簇，即能给出每个特征向量隶属于某个类簇的隶属度，即使对于很难明显分类的向量，Fuzzy-Kmeans 聚类也能得到较为满意的效果。相较于其他 4 种聚类方法，Fuzzy-Kmeans 聚类更适合文本聚类。Mahout 工具提供了 Fuzzy-Kmeans 的串行及并行实现，可以很好地解决海量文本的词聚类问题。

为了得到质量较高的词类簇，本文在使用 Fuzzy-Kmeans 算法进行聚类分析的同时，辅以人工整理文本训练集、调整并确认合理的聚类阈值。在分布式计算环境下，词聚类的并行实现包括如下几个关键步骤：

(1) 序列化文本：Mahout seqdirectory -I 源文本 -o 序列化文本 -ow

(2) 向量化文本：Mahout seq2sparse -i 序列化文本 -o 向量化文本-lnorm -nv -wt tfidf

向量化后的文本的目录结构如下：df-count 保存文本频率信息；tf-vectors 保存以 TF 作为权值的文本向量信息；tfidf-vectors 保存以 TFIDF 作为权值的文本向量信息；tokenized-documents 保存分词过后的文本信息；wordcount 保存全局的词频；dictionary.file-0 保存文本的词汇表；frequcency-file-0 保存词汇表对应的频率信息。文本聚类时将用到 tfidf-vectors 数据集，在查看聚类结果时将用到 dictionary.file-0。

(3) 聚类：Mahout fkmeans -i 向量化文本/tfidf-vectors -c /fkmeans-clusters -o /聚类结果 -k 20 -dm org.apache.Mahout.common.distance.CosineDistanceMeasure -m 1.05 -x 200 -ow --clustering --convergenceDelta 0.01

其中：-i：使用向量化文本的 TFIDF 向量作为输入；-o：输出聚类结果，包括每一次迭代后的聚类结果；-k：设定类簇个数；-c：聚类中心点。若不设定 k，则用这个目录里面的点，作为聚类中心点。否则，随机选择 k 个点，作为中心点；-dm：距离公式，文本聚类推荐用 cosine 距离；-m：归一化系数，须大于 1；-x：最大迭代次数；--clustering：mapreduce 并行模式；--convergenceDelta：迭代收敛阈值，默认 0.5。

(4) 提取每个类簇中权重最高的 N 个特征词：Mahout clusterdump -i /聚类结果/clusters-n-final -d /dictionary.file-0 -dt sequencefile -o/结果输出 -n 20 -b 50

其中：-i：最终迭代生成的簇结果；-d：从文本的词汇表中查找“词与词 id”的映射，使得输出结果中，可以直接显示每个类簇的权重最高词文本，而不是词 ID；-dt：映射类型；-o：结果输出目录；-n：提取前 N 个权重最高的词；-b：每行提取前 n 个字符。

(5) 将表示类簇的特征词及其 TFIDF 权重信息保存到 RMDB 数据库中。

3.3 基于词聚类的文本分类

基于词聚类的文本分类，可以看作是词类簇和测试文本之间的相似度计算，并依据计算结果，将测试文本的类别划归到相似度最高的几个词类簇中。目前存在多种基于 VSM 的文档相似度算法，如贝叶斯算法、神经网络算法等。本文采用效率高且易于实现的“简单向量距离法”[16]，其相似度计算的精度能满足一般性的要求，另外，该算法规范了文本特征向量的长度，这意味着在计算相似度时，不会放大数据对象重要部分的作用[17]。此算法的实现思路是将文档映射成向量空间中的点，点之间的距离用向量间的余弦夹角

来度量，即表示了文档间的相似程度。本文假设 Di 为词类簇，Dj 为测试文本，测试文本需经过特征向量筛选，之后表示 Di、Dj 的向量分别为：A = f(di)，B=f(dj)。这里将 A 和 B 分别记为：A = {A1, …, Ai, …, An}，B ={B1, …, Bi, …, Bn}，那么，根据夹角余弦定理，词类簇与测试文本的相似度为标准向量点积除以两个向量的长度积。

$$sim(di,dj) = cos(di, dj) = \frac{\sum_{k=1}^{n} d_{ik} \times d_{jk}}{\sqrt{(\sum_{k=1}^{n} d_{ik}^2)(\sum_{k=1}^{n} d_{jk}^2)}} \quad (1)$$

其中，n 为向量维数，d_k 为词语在向量集合中的第 k 维权值。由公式(1)可知，词类簇与测试文本 Di 和 Dj 的词特征向量 A 和 B 的夹角余弦值越接近 1，说明其向量间距离越短，即词类簇与测试文本之间的距离越短，相似度越大。测试文本与词类簇相似度计算的主要实现过程及伪代码如下：

```
数据库连接器;
提取测试文本;
  使用 IK 分词器对测试文本分词，得到文本特征向量;
提取词类簇信息 grs;
 While (类簇信息 grs.next){
      通过类簇信息提取该词类簇的特征向量 trs;
 While (特征向量 trs.next){
      If (特征向量也是文本向量，即公共子向量){
           提取该文本向量词频及文本向量长度,
                计算该文本向量 TF 权重;
           提取相应词类簇向量 TFIDF 权重;
      }
 }
利用相似度计算公式，计算测试文本与该类簇的相似度;
}
相似度排序，得到推荐类簇的结果。
```

4　分布式环境下的文本聚类开发运行环境

根据分布式环境下的文本聚类计算总体策略和思路，充分利用 Hadoop、Mahout 和关系型数据库在处理海量数据时的优缺点[18]，本文将需要大数据存储和数据挖掘计算的部分交给 Hadoop 及 Mahout 平台完成，将需要较强表达能力的查询交互部分交给关系型数据库完成，形成可靠性高及数据处理能力强的大规模计算系统环境。整个系统运行环境主要包括如下 5 个部分。

(1) 开源分布式环境(Hadoop[13])，考虑到稳定性和兼容性，选用 Hadoop2.2.0 版本。

(2) 分布式数据挖掘平台(Mahout)，本文选用 Mahout-distribution-0.9 版本。

(3) 数据处理平台(PostgreSQL[19])，在分布式文本聚类计算框架下，PostgreSQL 主要用于文本分类和词聚类的结果存储与查询。

(4) 开发平台(Eclipse[20])。

(5) Web 应用平台(Tomcat[21])。

分布式环境下的文本聚类开发运行环境涉及的平台和参数较多，在 etc/profile 中，除了 Java 相关的环境变量外，还需要准确配置 Hadoop、Mahout、PostgreSQL 环境变量保证整个计算平台的稳定正确运行。

5　实证研究

结合以上算法，本文给出相应实验数据。实验中使用的数据集是20Newsgroups[22]数据集，该数据集是卡内基·梅隆大学的 Lang 于1995年收集并整理的包含19 997篇文档约平均分布在20个类别中的 Usenet 新闻组语料，文章大小在1K-50K 之间，平均3K。Newsgroups 已经成为文本分类及聚类应用及测试中常用的数据集。实验环境为使用5台虚拟服务器形成 Hadoop 集群，一台主节点(Namenode)，一台辅节点(SecondaryNameNode)，三台数据节点(Datanode)；虚拟服务器配置为 Intel (R) Core (TM)2 Duo CPU E8400 @ 3.00GHz, 2GB 的内存，操作系统为 CentOS Release 6.5 i586。基于词聚类的文本聚类计算中，词聚类的质量对整个计算效果起到了至关重要的作用，因此，并行环境下，迭代收敛阈值和训练文本集大小对词聚类质量的影响是本文主要测试指标。

5.1　迭代收敛阈值

Mahout 聚类算法的迭代收敛阈值默认为 0.5，并不适合文本聚类。这里测试[0.006, 0.06]之间的阈值参数，使用 20Newsgroups 数据集的 MINI 集，共 2 000 个文本，每组(类簇)100 个，每个类簇取前 10 个权重最高的词作为该类簇的识别特征，与 mini_20Newsgroups 的人工分类结果作对比并计算聚类准确率。从图 3、图 4 可以看出，迭代收敛阈值为 0.01 时，聚类准确率较高，同时迭代时间(分钟 M)、次数较低，聚类效果较好。

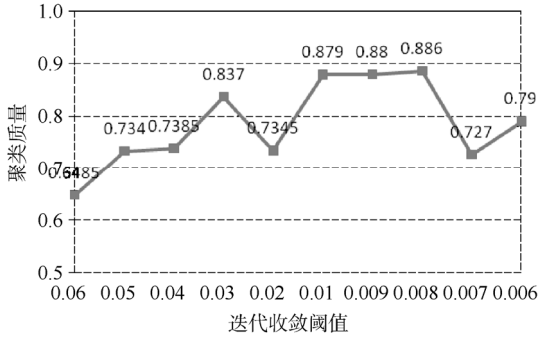

图 3　文本聚类准确率与迭代收敛阈值关系

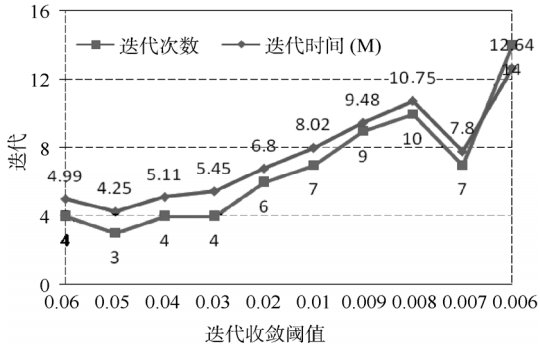

图 4　迭代收敛阈值与迭代次数及时间的关系

5.2　训练文本集规模与聚类质量

并行环境下，较大的训练文本集能够得到质量更好更稳定的特征向量集合，可以更好地挖掘和揭示隐藏在数据下面的关联关系。本文在原始数据集的 20 个 Newsgroups 组中每组分别随机提取 20、40、60、80、100、120 个文本为基础训练集，测试文本集大小和聚类质量之间的关系，从图 5、图 6 中可以看到当每组的文本个数分别超过 40 和 100 时，即特征向量个数趋于收敛时，文本聚类质量有较明显的提升。从图 6、图 7 中可以看出聚类迭代次数及时间与测试文本集大小成正相关关系，迭代次数增幅不大，迭代时间增幅略大于迭代次数的增幅。

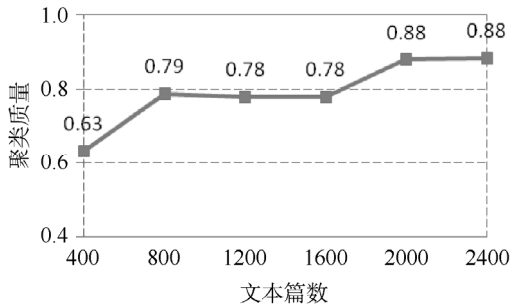

图 5　文本集规模与聚类质量关系

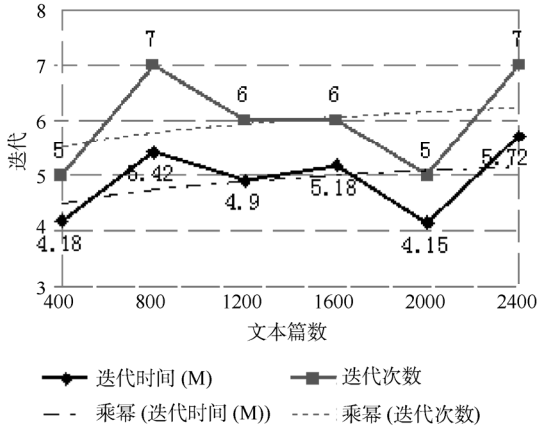

图 6　文本集规模与聚类次数及时间关系(1)

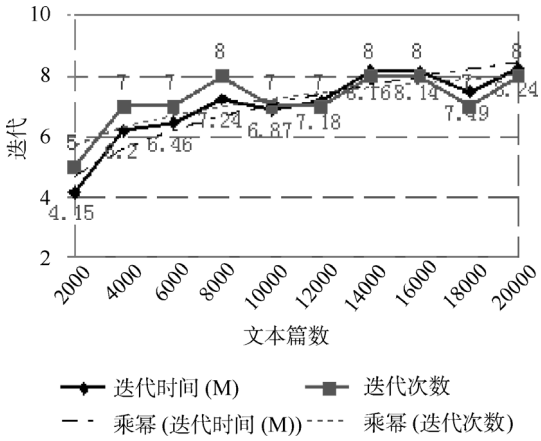

图 7　文本集规模与聚类次数及时间关系(2)

6　结　语

本文利用开源大数据挖掘工具，通过基于词聚类的文本聚类分类并行化计算方法，有效解决海量文本的词聚类瓶颈问题，结合开源软件及开发框架，提升文本聚类分类质量和效率，并给出搭建基于词聚类的文本聚类分类算法的开发环境构架和关键步骤，使读者能够明确掌握 Hadoop 平台实质、Mahout 工具的使用及如何实现分布式并行算法的快速部署解决方案。通过实验对比，本文遴选得出适合文本聚类的迭代收敛阈值和训练文本集规模。数据预处理优化及领域相关的词聚类优化是整个文本聚类分类质量的重点，也是下一步的工作方向。

参考文献：

[1]　胡建军，唐常杰，李川，等．基于最近邻优先的高效聚类算法 [J]．四川大学学报：工程科学版，2004，36(6)：93-99.

(Hu jianjun, Tang Changjie, Li Chuan, et al. An Efficient Multi-layer Clustering Algorithm Based on Nearest Neighbors First [J]. Journal of Sichuan University: Engineering Science Edition, 2004, 36(6): 93-99.)

[2] Han J, Kamber M. Data Mining Concepts and Techniques [M]. Beijing: China Machine Press, 2008: 261-284.

[3] Pena J M, Lozano J A, Larranaga P. An Empirical Comparison of Four Initialization Methods for the K-means Algorithm [J]. Pattern Recognition Letters, 1999, 20(10): 1027-1040.

[4] Bradley P S, Fayyad U M. Refining Initial Points for K-means Clustering [C]. In: Proceedings of the 15th International Conference on Machine Learning (ICML'98). San Francisco, USA: Morgan Kaufmann Publishers Inc., 1998: 91-99.

[5] Steinbach M, Karypis G, Kumar V. A Comparison of Document Clustering Techniques [C]. In: Proceedings of KDD 2000 Workshop on Text Mining. 2000: 1-20.

[6] Zhao Y, Karypis G, Fayyad U. Hierarchical Clustering Algorithms for Document Datasets [J]. Data Mining and Knowledge Discovery, 2005, 10(2): 141-168.

[7] Higgs R E, Bemis K G, Watson I A, et al. Experimental Designs for Selecting Molecules from Large Chemical Databases [J]. Journal of Chemical Information and Computer Sciences, 1997, 37(5): 861-870.

[8] Snarey M, Terrett N K, Willet P, et al. Comparison of Algorithms for Dissimilarity-based Compound Selection [J]. Journal of Molecular Graphics & Modelling, 1997, 15(6): 372-385.

[9] Slonim N, Tishby N. Document Clustering Using Word Clusters via the Information Bottleneck Method [C]. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'00). New York, USA: ACM, 2000: 208-215.

[10] 梁维铿. 基于 Hadoop 的分布式文本聚类研究[D]. 广州: 华南理工大学, 2011. (Liang Weikeng. Research of Distributed Text Clustering Basic on Hadoop [D]. Guangzhou: South China University of Technology, 2011.)

[11] MapReduce [EB/OL]. [2014-08-06]. http://Hadoop.apache.org/mapreduce/.

[12] Mahout [EB/OL]. [2014-08-06]. http: //mahout.apache.org/.

[13] Hadoop [EB/OL]. [2014-08-06]. http: //hadoop.apache.org/.

[14] Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing [J]. Communications of the ACM, 1975, 18(11): 613-620.

[15] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms [M]. Springer US, 1981.

[16] Salton G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer [M]. Boston: Addison-Wesley Longman Publishing Co., Inc., 1989.

[17] 田润涛, 谢培山. 色谱指纹图谱相似度评价方法的规范化研究(一) [J]. 中药新药与临床药理, 2006, 17(1): 40-42. (Tian Runtao, Xie Peishan. Study on the Standardization of Similarity Evaluation Method of Chromatographic Fingerprints (Part I) [J]. Traditional Chinese Drug Research & Clinical Pharmacology, 2006, 17(1): 40-42.)

[18] Pavlo A, Paulson E, Rasin A, et al. A Comparison of Approaches to Large-scale Data Analysis [C]. In: Proceedings of the 35th SIGMOD International Conference on Management of Data (SIGMOD'09). New York, USA: ACM, 2009: 165-178.

[19] PostgreSQL [EB/OL]. [2014-08-06]. http://www.postgresql.org/.

[20] Eclipse [EB/OL]. [2014-08-06]. http://www.eclipse.org/.

[21] Apache Tomcat [EB/OL]. [2014-08-06]. http://tomcat.apache.org/.

[22] Joachims T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization [C]. In: Proceedings of the 14th International Conference on Machine Learning (ICML'97). San Francisco, USA: Morgan Kaufmann Publishers Inc., 1997: 143-151.

收稿日期: 2014-07-14
收修改稿日期: 2014-08-22

Research and Implementation of Textual Clustering in Distributed Environment

Zhao Huaming
(National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: [**Objective**] To implement the textual clustering and classification in distributed environment through open-source tools. [**Methods**] According to the convergence of words in masses of text, this paper classifies texts based on word-clustering, including text preprocess by open-source tokenizer, cluster analysis by Mahout, classifying the test text by computing the similarity between the text and word-cluster. [**Results**] The textual clustering based on word-clustering in distributed environment effectively solves the bottleneck of word-clustering of massive texts. The tested result of word-clustering is ideal while the number of text training set exceeds 100 and the iterative convergence threshold is 0.01. [**Limitations**] The data type is limited in the field of news and the other field-based word-clustering also needs further test, optimization and adjustment. [**Conclusions**] This study describes the build process and key steps of the textual clustering and classification in distributed environment to help readers with in-depth understood.

Keywords: Distributed environment　Clustering　Textual clustering　Hadoop　Mahout

南京大学组建江苏省“数据工程与知识服务”重点实验室

在江苏省教育厅公布的2014年省高校重点实验室名单中，由南京大学信息管理学院牵头申报的“江苏省数据工程与知识服务重点实验室”获批立项，成为南京大学社会科学领域首个江苏省高校重点实验室。该实验室将依托国家重点学科情报学，联合江苏省科学技术情报研究所，集中申请单位和共建单位的研发优势，开展协同创新，解决大数据及相关服务领域的技术和应用问题，推动和引导江苏省大数据产业和科技服务领域的快速发展。

数据工程与知识服务重点实验室建设期为三年，欧洲文理科学院院士、南京大学信息管理学院叶鹰教授担任实验室主任，教育部长江学者、南京大学信息管理学院苏新宁教授为实验室学术带头人，并担任学术委员会主任，实验室专兼职研究人员近 50 人。实验室的主要研究方向为大数据环境下的数据整合与规划、知识关联技术、数据分析理论、知识服务技术等，将重点突出大数据环境下的知识服务，实验室将联合有关企事业单位，建立示范性相关数据平台、技术平台、服务平台，促进有关应用和技术的转化和推广。

数据工程与知识服务重点实验室为开放实验室，每年还将提供数据和设备平台，设立开放课题供国内外广大学者共同参与数据工程与知识服务的研究。2015 年度开放课题研究方向如下，欢迎全国广大学者积极申报。

(1) 大数据整合与规划：政策与战略
(2) 大数据知识关联：技术与应用
(3) 大数据分析：理论、模型与应用
(4) 大数据的行业影响：金融、医疗等
(5) 数据开放、数据交易与隐私保护
(6) 数据工程、数据资源融合

主页：http://deks.nju.edu.cn
电话：86-25-89683597
地址：江苏省南京市栖霞区仙林大道 163 号南京大学信息管理学院
邮编：210023

(本刊讯)