基于DK算法的互联网热点主动发现研究与实现

李若鹏,李 翔,林 祥,李建华
(上海交通大学 信息安全工程学院,上海 200240)

摘 要:针对互联网舆情管控领域信息量大,时效性强,往往偏重于某些方向,如社会热点、焦点,或反动、黄色言论等的特点,文中把基于密度的聚类思想引入传统K-Means算法,提出全新的DK聚类算法,并且基于DK算法构建中文文本聚类模型,重点对互联网媒体发布信息进行主动热点发现研究。用实验验证中文聚类模型的具体性能,证实了该模型的有效性和实用性。

关键词:K-Means;DK;中文文本聚类;舆情管控
中图分类号:TP391.3 **文献标识码**:A **文章编号**:1673-629X(2008)09-0001-04

Discovering Information Hotspots on Initiative over Internet Based on DK Clustering Algorithm

LI Ruo-peng, LI Xiang, LIN Xiang, LI Jian-hua
(School of Information Security, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: In the information booming era, Internet information control and supervision always need to deal with numerous update information and focuse on some specific areas such as social focus, hot topics, anti-social statement and porno information. Considering all these features, create a Chinese text clustering model and specialized in Internet information hotspots discovery on initiative. It proposes the density based DK solution also combined the strength of K-Means algorithm and the feasibility is justified in the experiment.

Key words: K-Means; DK; Chinese text cluster; information control and supervision

0 引 言

随着互联网的快速发展,通过互联网发布、检索信息已经成为越来越多的网络用户日常生活的重要组成部分,互联网上存储和传输的信息能够很大程度反映一定时期社会各领域人们所关注的热点、焦点。因此,深入分析研究互联网舆情管控领域中信息处理技术,尤其是热点信息主动发现技术,已经成为一项紧迫而又重要的课题。

传统信息处理技术对于文本基于内容的分析,主要包括文本分类技术和文本聚类技术。这两类技术的目的都是将大规模的文本数据对象分组形成多个类别,使得属于同一类的文本信息之间具有较高的相似度,而位于不同类别中的文本信息差别明显,从而方便网络用户对于文本信息的有效利用。但是,文本分类是已知一批训练文本的标签,通过机器学习得到文本分类器[1],它需要大量训练样本作为先验类别知识,因此不能适时反映互联网热点信息内容新、变化快和新类别层出不穷的特点。而文本聚类无需训练样本,所划分的类是未知的,能够应用于面向一段时间内有代表性的网络文本,主动发现该阶段互联网的热点和焦点。笔者正是将传统文本聚类算法加以改进后,构建新型文本聚类系统,应用于互联网热点信息主动发现,并通过系列实验验证改进技术的有效性和实用性。

收稿日期:2007-12-06
基金项目:上海市科委"登山行动计划"信息技术领域重点项目(065115020);国家自然科学基金项目(60502032)
作者简介:李若鹏(1982-),男,硕士研究生,研究方向为计算机网络应用层、内容安全及数据挖掘、文本聚类;李 翔,博士,副教授;李建华,教授,博导。

1 研究现状

传统数据挖掘研究派生出大量成熟的信息聚类算法,主要可分为划分方法(partitioning method),层次方法(hierarchical method),基于密度的方法(density-based method),基于网格的方法(grid-based method)和基于模型的方法(model-based method)五大类[2]。每一大类信息聚类算法都拥有代表算法,实际应用中聚类算法选择主要取决于数据类型、聚类目的和算法应用等。下面就互联网热点信息主动发现这一应用需求,对当前文本聚类中常用的、有代表性的聚类方法进

行分析,考察其在热点主动发现方面的有效性。

1.1 K-Means法

K-Means算法是所有聚类算法中应用最广泛的一种分割方法。它的基本思想为:给定一个例子集合 n 和一个整数 k ,K-Means算法将 n 分割为 k 个聚类并使得在每个聚类中所有值与该聚类中心的距离总和最小,其中聚类中心是指该聚类的几何平均值。该算法具有迭代速度快、能有效处理大数据集的优点。但是该算法存在一个显著的缺点,即需要事先给出 k 值,它经常中止于一个局部最优解[3],其聚类效果很大程度上取决于初始聚类中心的选取,这对于分类数无法预先获知,互联网文本信息内容很不规则导致很难选取合适的初始聚类中心点来说,显然是个严重的问题。

1.2 CURE法

它不用单个质心或对象来代表一个簇,而是选择数据空间中固定数目的具有代表性的点。一个簇的代表点通过如下方式产生:首先选择簇中分散的对象,然后根据一个特定的分数或收缩因子向簇中心"收缩"或移动它们。在算法的每一步,有最近距离的代表点对(每个点来自于一个不同的簇)的两个簇被合并。

优点:可适应非球形的几何形状,簇的收缩或凝聚有助于控制噪声的影响,效率高。

缺点:要求用户给出一些参数,例如样本大小、希望聚类的数目及收缩因子,这些参数都是难以确定的,同样不适用于文中的研究背景。

1.3 DBSCAN法

DBSCAN是一个基于密度的聚类算法,它是在数据集上定义一种密度可达等价关系,对应的划分就是聚类。算法思想是:检查一个对象的 e 领域的密度是否足够高,即一定距离 e 内数据点的个数是否超过Minpts,依此确定是否建立一个以该对象为核心对象的新簇,再合并密度可达簇。该算法能在带有"噪声"的空间数据库中发现任意形状的聚类,但必须输入参数 e 和Minpts,且聚类结果对这两参数比较敏感,另外算法的时间复杂度偏高,难以适应大规模文本信息的聚类要求。

1.4 SOM法

这是一种无监督的聚类方法,通过反复学习来聚类数据,其聚类过程是通过若干个单元竞争当前对象来进行的;为了更接近输入对象,对获胜单元及其最近邻居的权重进行调整。

优点:可视化、拓扑结构保持以及概率保持。

缺点:当学习模式较少时,网络的聚类效果取决于输入模式的先后顺序,而且网络连接权向量的初始状态对网络的收敛性能有很大影响。

综上,现有的传统聚类方法不能直接应用于文中的研究背景。为了有效提高发现海量未知信息的速度和准确率,必须找到一种速度快,对先验知识要求尽可能少的聚类算法。

2 DK算法实现原理

K-Means聚类算法具有迭代速度快、能有效处理大数据集的特点,但无法解决初始聚类中心的选取问题,鉴于此,笔者引入密度聚类的思想,提出了基于密度与K-Means相结合的DK聚类算法。

DK算法,即Density-kmeans算法,该算法可自动发现聚类类别数并且选择有效的初始聚类中心点,克服了K-Means算法需要事先给出 k 值、初始聚类中心选取困难、对孤立点和噪声数据很敏感的不足,同时克服了基于密度聚类算法速度慢的缺点,操作步骤:

1) 将每个文本向量都看成一个独立类别,计算所有向量之间的距离,生成距离矩阵;

2) 选取2个正数,一般 $R2 = 2R1$,其中 $R1$ 为所有向量之间距离的平均值;

3) 以每个向量为球心,以 $R1$ 为半径作球,计算落在每个球内的向量数目,即样本密度;

4) 将样本密度按从大到小的顺序排列,取密度最大者作为第一个凝聚点Z1,在密度次大的单元中任选一点 k,若与第一凝聚点之间距离大于 $R2$,即 $|Z1-k|>R2$,则把 k 作为第二个凝聚点Z2,否则继续判定下一密度最大者,若下一密度最大者中的任一点与前面若干个凝聚点之间距离均大于 $R2$,则将之作为又一新的凝聚点,如此反复迭代直到没有新的凝聚点生成;

5) 这些凝聚点作为聚类模板的初值即分类个数 k 以及初始 k 个聚类中心 $Z1,Z2,Z3,\cdots,Zk$;

6) 把得到的 k 和 k 个聚类中心 $Z1,Z2,Z3,\cdots,Zk$ 作为K-Means算法的初始模板,继续用K-Means算法迭代,最后得到 k 个聚类;

7) 经过初始聚类,可以得到全部向量的聚类个数 k,以及模板初始聚类中心 $Z=\{Z1,Z2,Z3,\cdots,ZR\}$,然后,进行K-Means迭代,使得每个文本向量根据与聚类中心距离的远近程度,形成 k 个互不相交的聚类,较为相似的向量都聚在同一类中。

改进后的聚类算法是动态的,初始模板的设定不再依赖人的经验参数,而是从整个文本向量空间的统计特性中获取必要的参数信息,因而最后的聚类结果更加客观合理。

3 基于DK算法的中文文本聚类模型

基于创新性提出的DK算法,设计实现了一个中

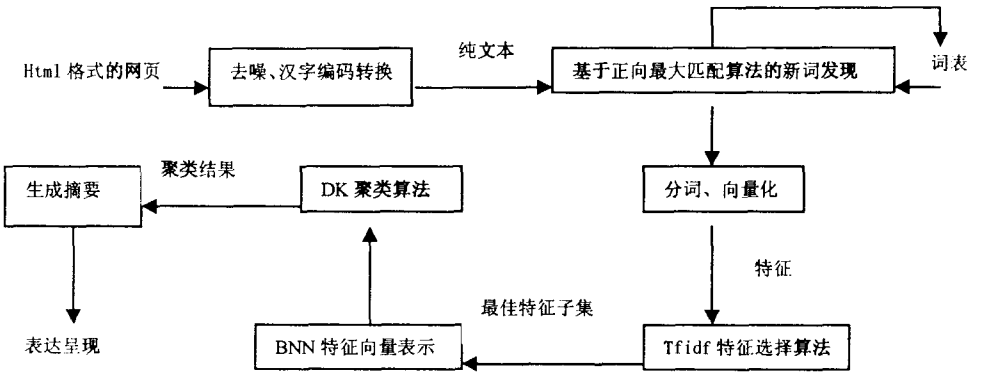

图1　中文文本聚类模型(CTCM)

文文本聚类系统(见图1),主要包括新词发现、特征表示、聚类算法以及基于聚类的类描述信息自动生成等模块。

首先,在新词发现模块,使用了基于正向最大匹配算法的任意长特征值发现机制,可以实现动态的调整词表,跟踪网络最新热点信息。

其次,在特征选择和特征向量表示模块,采用了tfidf算法进行特征选择,增加了基于BNN的特征向量表示。其中,tfidf方法是信息获取研究中经常使用的一种加权词频算法。它基于两个直觉性的认识:特征在文本中出现次数越多,越重要;特征在越多的文本中出现,越不重要[4]。对于任意给定的单词w,其tfidf值的计算公式表示如下:

$tfidf(t) = tf(t) \times idf(t)$, 其中 $idf(t) = \log(n/n(t))$。

$tf(t)$是特征值t在所有文本中出现的次数,$n(t)$是出现特征值t的文本数,n是全部文本数。

具体步骤是先用正向最大匹配法找出文本中出现的特征词,统计该特征词在所有文本中出现的次数,然后用tfidf算法计算出经过加权的值,取最大的前N个特征值作为N维向量空间。对于选出的N个特征值,采用bnn算法进行特征向量表示,即只要文本中出现过该特征词就在对应维上置1,否则置0。

向量间相似度的表示采用两向量之间的夹角余弦[5],即:

$$sim(d_k, c_j) = \frac{V(d_k) \cdot V(c_j)}{|V(d_k)| \times |V(c_j)|}$$

再次,在聚类算法模块,通过对现有聚类算法的仔细分析和比较,结合K-Means划分方法和基于密度的聚类算法,给出一个适合于大规模数据的、快速高效的,并且无需设定初始聚类中心,可自动发现聚类数的递增聚类方法——DK算法,并通过试验结果分析该算法在舆情管控热点信息领域中的优劣,对算法性能和查全率、查准率进行具体分析。

最后,自动生成每一个类的描述信息,包括该类包含的热点词及中心段落等。

实验证明,该系统具有良好的稳定性和聚类准确性,对于互联网热点信息的获取与应用效果显著。

4　实验描述与测试结果

本系统实验环境为Windows平台,VC6.0编译器。测试样本为sogou语料库的精简版本9个类近18000篇文本,使用tfidf算法进行特征向量选择。

测试内容包括:

对于K-Means算法和DK算法,从一个样本类(如经济类)中选取一定数量的目标文本与2000篇杂类文本混合组成一组实验样本,在不同目标文本数和不同维度的向量空间中(可选取400维和1200维),测试每一组实验样本的查全率和查准率。

首先给出查全率与查准率的定义:

查准率,是指聚类判定的属于类别C的所有文档中,确实属于类别C的文档所占的比例;查全率,是指原本属于类别C的所有文档中,聚类做出同样判定的文档所占的比例。

具体步骤:

首先,依次随机抽取财经(08)、体育(14)、军事(24)类各500,800,1200,1600,2000篇目标文本与2000篇杂类文本混合组成三组实验样本。

其次,分别使用K-Means算法和DK算法,测试在400维向量空间中每一组实验样本在不同目标文本数条件下的查全率(见图2,3)和查准率。

再次,分别使用K-Means算法和DK算法,测试在1200维向量空间中每一组实验样本在不同目标文本数条件下的查全率和查准率(见图4,图5)。

最后,进行实验结果的对比与分析:

1)在向量空间维度一定的条件下,采用基于密度改进的K-Means算法-DK算法,其聚类准确度得到了显著提高。

2)在聚类算法和向量空间维度一定的条件下,选

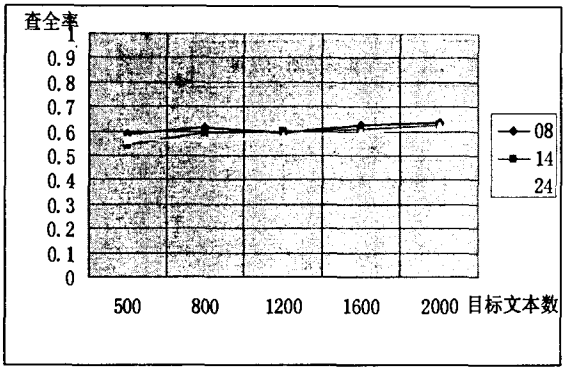

图 2 K－Means 聚类算法查全率 400 维

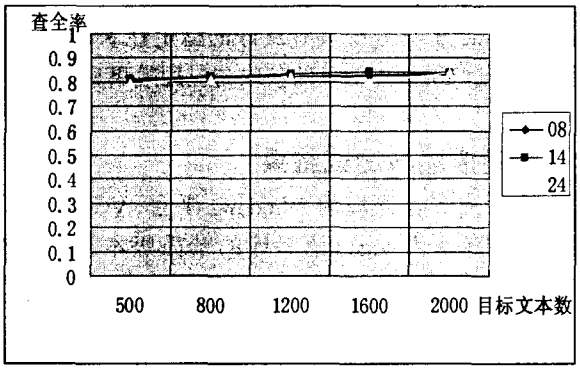

图 3 DK 聚类算法查全率 400 维

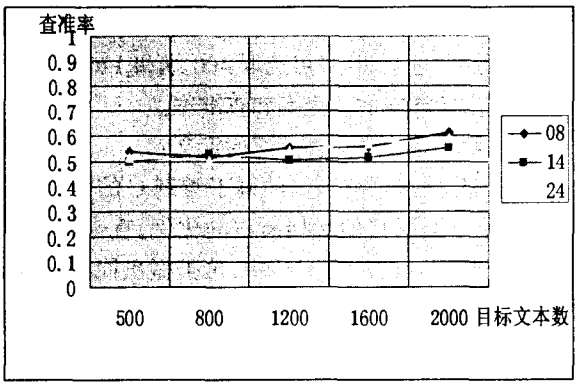

图 4 K－Means 聚类算法查准率 1200 维

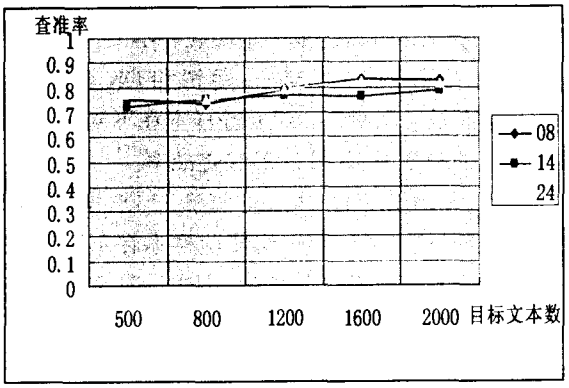

图 5 DK 聚类算法查准率 1200 维

取的目标文本数量越多，聚类准确率和稳定性会有所提高。这是因为目标文本数量越多代表目标类文本的向量空间特征值越多，该类的独立性也越强。当目标文本数量足以提供该类代表的信息之后，聚类查全率和查准率趋于稳定。

3) 在聚类算法和目标文本数量一定的条件下，1200 维向量空间比 400 维向量空间的聚类准确率略高。这是因为在海量文本条件下，1200 维向量空间包含了更多的类特征信息，更能有效地区分类之间的差别，但是并不是维数越大越好，由于特征向量的维数是 tfidf 值最高的 N 个特征词，当 N 超过代表每个类独有的特征信息总和后，会引入大量冗余词，导致聚类准确率下降。

4) 采用 DK 算法，聚类查全率和查准率最终接近 80%，不再有所提高。这是因为一篇文本是否属于一个类很大程度受先验知识的影响，单从特征向量之间的距离分析，同属一个类的文本不一定具有最近的距离，可能分到不同的类中。

5 结束语

笔者针对互联网舆情信息管控领域的热点信息主动发现，设计了一个中文文本聚类系统，通过对各种聚类算法的分析、比较与实验，提出了基于密度与 K－Means 相结合的聚类算法—DK 算法，克服了单纯采用 K－Means 算法依赖初始聚类数和初始聚类中心点的缺陷。大量的实验数据显示，DK 算法显著提高了聚类准确率。文中提出的中文文本聚类系统可满足海量文本数据的自动聚类，也可作为进一步分类的基础。

本系统可以加入自然语言理解的知识，对中文词汇的词性、词义进行分析，例如给动词和名次赋予高权重，淘汰无用的连词、助词，加入词的相关性分析等。这样可以提高向量空间的信息量，进一步改善聚类准确率。

参考文献：

[1] 李家福，张亚非，陆建江. 模糊聚类算法在汉语文本聚类中的应用[J]. 计算机工程，2002，28(4)：15－16.

[2] Han Jiawei，Kamber M. Data Mining Concepts and Techniques － 数据挖掘概念与技术[M]. 范 明，孟小峰等译. 北京：机械工业出版社，2006.

[3] 张 蓉. 数据聚类技术的研究[J]. 计算机工程与应用，2002(16)：145－147.

[4] 薛德军. 中文文本自动分类中的关键问题研究[D]. 北京：清华大学，2004.

[5] 郭庚麒. Web 文本挖掘技术[J]. 计算机与网络，2004(1，2)：114－116.