

互联网媒体信息热点主动发现技术研究与应用

黄宇栋, 李翔, 林祥

(上海交通大学信息安全工程学院, 上海 200240)

摘要:网络媒体庞大的信息量及信息内容的各异,并不能把传媒聚类算法中适用于文本信息聚类的基本划分方法直接应用于互联网媒体信息热点主动发现的研究工作中。鉴于此,文中将基于密度的聚类思想引入CFK-Means算法,创造性地提出了全新的DCFK聚类算法。与此同时,文中基于DCFK算法构造大规模中文信息聚类模型,并且通过系列实验验证本聚类模型在互联网媒体信息主动热点发现领域的有效性和实用性。

关键词:DCFK;大规模中文信息聚类;热点发现

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2009)05-0001-04

Research and Application on Active Discovery Technique of Internet Media Information Hotspots

HUANG Yu-dong, LI Xiang, LIN Xiang

(School of Information Security, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: Abundant information and difference of Internet media can't directly apply the basic divided method, which suit the text information cluster in media cluster algorithm, to actively discover the hotspots of Internet media information. So that, propose the the new cluster algorithm called DCFK based on CFK-Means algorithm. It is necessary to construct abundant Chinese information cluster model to discover the Internet information hotspots. Finally, series of experiments will prove the effectiveness and utility of them.

Key words: DCFK; Chinese information cluster; hotspot discovery

0 引言

随着互联网的快速发展,通过互联网获取信息、发布信息已经成为人们生活的重要组成部分,互联网媒体新发布信息很大程度反映当前社会各领域所关注的热点、焦点。然而,在“网络摩尔定律”的支配下,互联网信息量正以每100天翻一番的速度增长。面对增长如此迅速的新网络信息,如何快速、有效地主动发现互联网热点信息已经成为一项紧迫而又重要的课题。

1 研究现状

当前互联网媒体信息热点已经由通过信息分析专家人工归纳、手动提炼生成,逐步转向利用计算机系统在不断推陈出新的海量网络信息中自动聚类得到,聚

类结果中包含巨量信息的类别正是需要寻找的互联网媒体信息热点^[1]。因此高效、实用的互联网信息热点主动发现系统需要拥有能够高速处理大规模数据的信息聚类算法,不过目前在面向互联网海量信息聚类方面,暂未有成熟的核心算法与解决方案公开发表^[2]。

传统的聚类方法主要包含基于划分方法、基于层次方法、基于密度方法、基于网格方法和基于模型方法五种^[3],每种聚类方法特点如下。

(1)基于划分方法具有迭代速度快、能有效处理大数据集的优点,广泛应用于文本聚类,但需事先确定K个初始聚类中心,K-means算法是典型的划分方法。

(2)以CURE算法为代表的基于层次方法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,但聚类质量依赖于参数选择,并且这些参数在大规模数据条件下难以确定,故不适用于大规模文本聚类。

(3)基于密度方法能在带有“噪声”的空间数据库中发现任意形状的聚类,无需事先确定初始聚类中心,适用于文本聚类,但算法的时间复杂度较高,不适合处

收稿日期:2008-08-29

基金项目:国家自然科学基金项目(60502032);上海市科委项目(065115020)

作者简介:黄宇栋(1983-),男,硕士研究生,研究方向为内容安全、文本聚类;李翔,副教授,博士,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。

理大规模数据。

(4) 基于网格方法提出的网格结构有利于并行处理和增量更新,效率高,但聚类质量取决于网格结构的最低粒度,而最低粒度在大规模数据条件下难以确定,很少应用于文本聚类。

(5) 基于模型方法的时间复杂度高,不适合处理大规模文本信息^[3]。

考虑到网络媒体信息数量庞大,聚类运算计算复杂度高;同时网络媒体信息内容各异,难以事先确定初始聚类中心,因此并不能把传统聚类算法中适用于文本信息聚类的基于划分方法和基于密度方法直接应用于互联网媒体信息热点主动发现的研究工作中。鉴于此,笔者在有学者把聚类特性引入传统 K-Means 算法,形成 CFK-Means 算法的基础上,进一步将基于密度的聚类思想融入 CFK-Means 算法,提出全新的 DCFK 聚类算法,并且基于 DCFK 算法构造中文聚类模型进行互联网媒体信息热点发现研究。文中最后通过系列实验,证明基于 DCFK 算法的中文聚类模型的有效性 with 实用性。

2 背景知识与核心算法

2.1 聚类特性的定义与性质

聚类特性^[4](Clustering Features, CF)定义为三元组:

$$CF = (N, LS, SS) \quad (1)$$

其中 N 是聚类类别包含文本向量数目; $LS = \sum_{i=1}^N x_i$ 是 N 个文本向量的线性和,仍为一个向量; $SS = \sum_{i=1}^N x_i' x_i$ 是 N 个文本向量的平方和, SS 为一个数值。聚类类别中心点可以由聚类特性直接计算得到,如式(2)所示。

$$\text{聚类中心点 } \bar{x} = LS/N \quad (2)$$

聚类特性可加性:设 A 和 B 为两个行将合并的子聚类,两个子类别聚类特性分别为 $CF_A = (N_A, LS_A, SS_A)$ 和 $CF_B = (N_B, LS_B, SS_B)$,则合并后的聚类 C 的聚类特性是:

$$CF_C = CF_A + CF_B = (N_A + N_B, LS_A + LS_B, SS_A + SS_B) \quad (3)$$

2.2 DCFK 算法介绍与过程实现

2.2.1 DCFK 算法介绍

最近有学者利用聚类特性的独特性质,结合传统的 K-Means 算法提出了高效聚类算法——CFK-Means 算法^[5]。CFK-Means 算法利用聚类特性的可加性和表示性,构造出一种可递增的、批处理的、适用

于大规模数据的聚类算法。CFK-Means 算法首先继承 K-Means 算法实现简单、迭代速度快的特点;其次它通过累加批量数据已经运算获得的子数据集聚类特性,分批处理大规模数据,有效降低算法计算复杂度和算法实现时间,减少大规模数据聚类运算对于计算机系统的资源需求^[4]。但是 CFK-Means 聚类算法与传统的 K-Means 算法一样,仍然无法解决聚类类别数 K 和初始聚类中心点选取困难的实际问题,这正是本文核心——DCFK 算法需要重点解决的难点。

考虑到 CFK-Means 算法难于进行聚类类别数 K 和初始聚类中心点的选取,文中进一步将基于密度的聚类思想引入 CFK-Means 算法,首创性地提出同时融合聚类特性与密度思想的 Density-CFK-Means (DCFK) 算法。DCFK 算法秉承 CFK-Means 算法批量处理大规模文本向量方案,首先通过基于密度的聚类思想,面向第一部分文本向量,确定聚类类别数与初始聚类中心。其后 DCFK 算法转入 CFK-Means 算法迭代环节,批量运算剩余文本向量,最终在整个数据集上完成信息聚类,实现信息热点主动发现。

2.2.2 DCFK 算法过程实现

DCFK 聚类算法源于传统 K-Means 算法^[6],同时融合聚类特性与基于密度思想,算法实现过程如下文所述。

(1) 利用基于密度的聚类思想确定聚类类别与初始聚类中心。

① 任意选取 M_1 (例如 1000) 个文本向量,分别计算两两向量之间的距离,生成一个 $M_1 \times M_1$ 距离矩阵;得到所有两两向量间距离的平均值 R_1 ,选取正数 $R_2 = 2R_1$ 。

② 分别以每个向量为球心, R_1 为半径作球,计算落在球内的文本向量数,得到 M_1 个样本密度。

③ 将 M_1 个样本密度按从大到小的顺序排列,选取样本密度最大的文本向量作为第一个凝聚点 Z_1 。继续在剩余向量中选取样本密度最大的文本向量 X ,若 X 与所有已生成的凝聚点之间距离均大于 R_2 ,即 $\forall m \in (1, 2, \dots, n), \exists |Z_m - X| > R_2$,则将 X 作为新的凝聚点 Z_{n+1} ,否则放弃。反复迭代直到没有新的凝聚点生成,最终得到 K ($K \leq M_1$) 个凝聚点 Z_1, Z_2, \dots, Z_k 。将这 K 个凝聚点作为 DCFK 算法第二环节——CFK-Means 算法迭代过程中的初始聚类中心。

(2) 利用 CFK-Means 算法在整个文本向量集上批量进行文本聚类运算。

① 利用基于密度思想得到的 K 个初始聚类中心,将第一部分 M_1 个文本向量进行 K-Means 聚类迭代,即以文本向量与聚类中心距离最短为原则,将 M_1 个

文本向量分别归到 K 个类中^[3]。重新计算聚类中心,若聚类中心发生变化,则利用新的聚类中心对 M_1 个文本向量重新进行 K -Means 迭代,迭代操作停止于所有聚类中心不再发生变化。各个类别包含的 N 个文本向量,与 N 个文本向量的线性与与平方和,共同构成对应类别的聚类特性,如式(1)所示。

② 选取 M_2 (例如 2000) 个文本向量与 ① 中得到的 K 个类别聚类中心向量进行 K -Means 聚类迭代, K 个类别聚类中心可以利用前次聚类运算生成的聚类特性根据式(2) 计算得到。在所有聚类中心不再发生变化后,利用聚类结果更新 K 个类别的聚类特性。在更新前,首先区分聚类特性与本次纳入该类别的文本向量。其次运算本次纳入该类别所有文本向量的聚类特性。最后,根据聚类特性可加性,如式(3) 所示将当前类别内所有聚类特性叠加,完成本次 K -Means 聚类迭代后各类聚类特性的更新。

③ 继续批量读取 M_2 个文本向量与前次聚类运算得到的 K 个类别聚类中心向量进行 K -Means 聚类迭代,迭代完成后更新 K 个类别的聚类特性。以此类推,直到对于所有文本向量完成信息聚类运算。

DCFK 聚类算法在初始聚类类别的选定过程中不再依赖于信息分析专家的历史经验,而是基于整个文本向量空间的密度特性,完成信息聚类类别的选择与确定。与此同时,DCFK 聚类算法还继承了 CFK-Means 算法批量处理大规模数据的优势,具备高效聚类海量信息能力,能够适用于互联网信息热点主动发现的研究工作。文中将进一步基于 DCFK 算法构造大规模中文信息聚类模型,并且利用系列实验证明该模型在互联网热点主动发现领域的有效性与实用性。

3 基于 DCFK 算法的大规模中文信息聚类模型

基于 DCFK 算法构造适用于互联网媒体信息热点主动发现的大规模中文信息聚类模型主要包含文本分词、特征选择、向量表示与核心聚类等功运算模块,如图 1 所示。

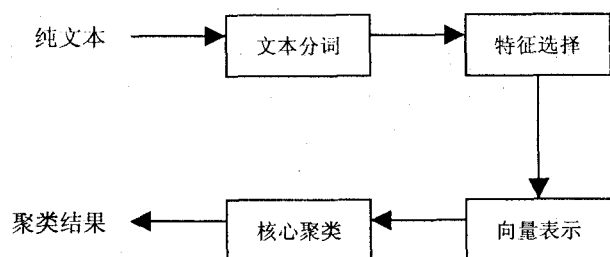


图 1 大规模中文信息聚类模型

* 文本分词模块。

文本分词模块以互联网媒体发布信息的纯文本内容为输入,输出是所有文本的分词结果。本模块进一步统计分词结果中的每个单词在整个文本向量集上出现的词频,以及包含该单词的文本数。本模块使用通用中文词库,基于前向最大匹配算法的任意长度特征词发现机制实现文本分词。

* 特征选择模块。

文本特征选择模块利用所有文本的分词结果,以 TFIDF 值最大的 N 个单词(参数 N 可根据聚类结果予以调整)共同构成的特征词库作为输出。其中,TFIDF 加权词频算法是信息处理领域中经常使用的特征选择算法。对于任意给定的单词 t ,TFIDF 值的计算方法如式(4) 所示。

$$\text{tfidf}(t) = \text{tf}(t) \times \text{idf}(t), \quad \text{其中 } \text{idf}(t) = \log(n/n(t)) \quad (4)$$

$\text{tf}(t)$ 是单词 t 在所有文本中出现的次数, $n(t)$ 是出现单词 t 的文本数, n 是全部文本数。

* 向量表示模块。

文本向量表示是将文本信息映射成文本向量的过程,文中采用 BNN 文本向量表示方法。首先把特征选择模块得到的 N 个特征词作为 N 维向量空间,每一维对应一个特征词;其次根据每个文本的分词结果,将文本信息映射成一个 N 维向量。如果文本中包含某个具体的特征词,就在该特征词对应的向量维度上置 1,否则置 0。利用 BNN 文本向量表示方法,把所有文本信息均映射为对应的 N 维文本向量。

* 核心聚类模块。

核心聚类模块正是利用文中提出的 DCFK 算法进行大规模中文信息的聚类运算。信息量大的类别就是需要寻找的信息热点,从而实现信息热点的主动发现。

文中后续章节将通过系列实验验证基于 DCFK 算法的大规模中文信息聚类模型,在互联网媒体信息热点主动发现工作中的有效性和实用性。

4 模型性能评估

为了检验基于 DCFK 算法的大规模中文信息聚类模型的有效性和实用性,文中对 CFK-Means 算法和 DCFK 算法进行一系列聚类运算对比实验,主要从聚类结果的查全率和查准率两方面进行模型性能评估。

4.1 性能评估环境

模型性能评估环境选择 Windows 平台,测试样本来源于国内重要网络媒体发布信息(含门户网站与 BBS 论坛)。经过手工分类和精心整理后,测试样本具体分为 9 大类别:财经、IT、健康、体育、旅游、教育、招聘、文化和军事。

4.2 性能评估过程

1) 任取语料库两大类, 每类依次随机抽取 500, 1000 篇目标文本, 总共产生四组目标文本。

2) 在语料库的其余七大类中, 任取 2000 篇文本作为杂类文本, 并分别与 1) 中选取的四组目标文本混合组成四组实验样本。

3) 分别使用 CFK - Means 算法和 DCFK 算法, 测试每一组实验样本在 128, 256, 512, 1024 维向量空间条件下的查全率和查准率, 查全率与查准率的定义如式(5)和式(6)所示。

查准率是指聚类判定属于类别 C 的所有文档中, 确实属于类别 C 的文档所占的比例。

查准率(Precision) =

$$\frac{\text{聚类后类别 C 中确实属于 C 的文本数}}{\text{聚类后类别 C 中的全部文本数}} \quad (5)$$

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。

查全率(Recall) =

$$\frac{\text{聚类前后都在类别 C 中的文本数}}{\text{聚类前原本在类别 C 中的文本数}} \quad (6)$$

4.3 性能评估结果与分析

论文选取财经、教育作为性能评估的两个大类, 每类分别随机抽取 500、1000 篇文本与 2000 篇杂类文本组成四组实验样本, 并且使用 CFK - Means 算法和 DCFK 算法, 分别测试每一组实验样本在 128, 256, 512, 1024 维向量空间中的聚类查全率和查准率, 如图 2~5 所示。

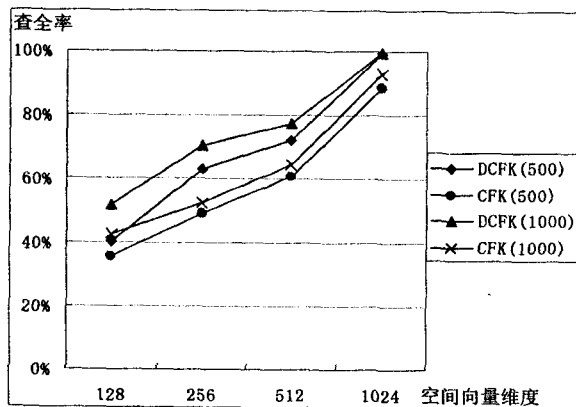


图 2 财经类查全率

根据 CFK - Means 算法和 DCFK 算法在不同向量空间中的聚类查全率和查准率的比较, 不难得出如下结论:

(1) 对比 DCFK 算法和 CFK 算法的中文聚类结果, 相同向量空间维度下 DCFK 算法对于不同数量的各组实验样本在聚类运算的查全率和查准率上均优于 CFK 算法。

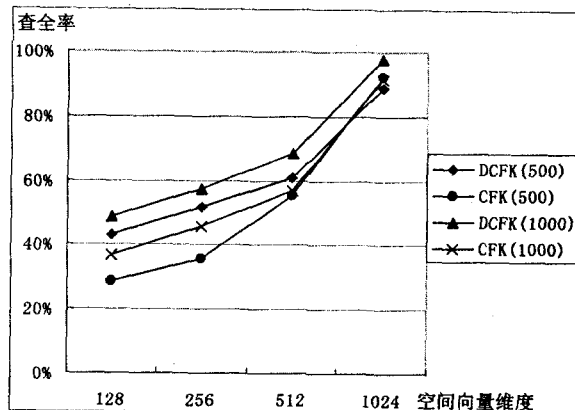


图 3 教育类查全率

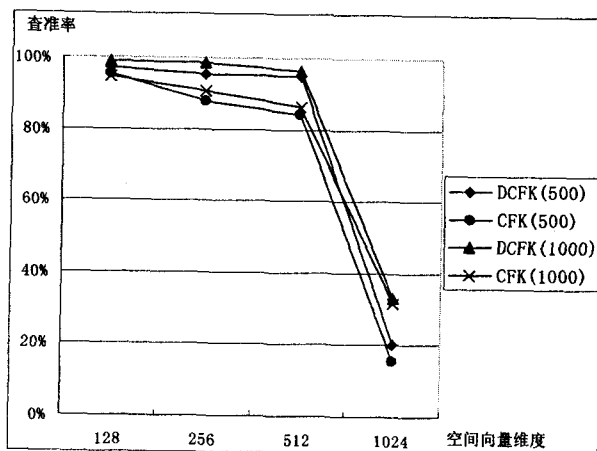


图 4 财经类查准率

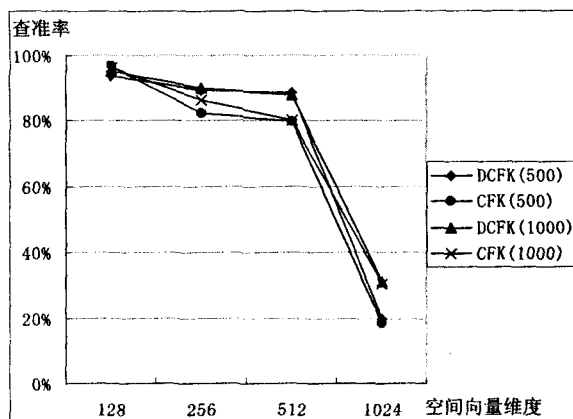


图 5 教育类查准率

(2) 在相同向量空间维度下, 增加正样本数量可以明显提高聚类算法的查全率和查准率。这是因为在包含负样本(杂类文本)的情况下, 正样本数量越多, 特征选择模块输出的特征词库越偏向正样本所属类别。当正样本数量足以提供所属类别代表信息后, 聚类运算的查全率和查准率趋于稳定。

(3) 在目标文本数选定的情况下, 适当增加空间向量维度可引入更多的关键词, 更有效地地区分不同类别之间的差异, 从而提高聚类运算的查全率和查准率。但维

(下转第 187 页)

3 可视化仿真结果与视图

图4显示的是文中三维可视化仿真中实现的模拟多颗卫星围绕地球运行的三维空天全景视场,红色轨道卫星为当前选取的关注卫星。

图5显示的是文中三维可视化仿真中实现的空天观测视场拉近后的卫星视场。

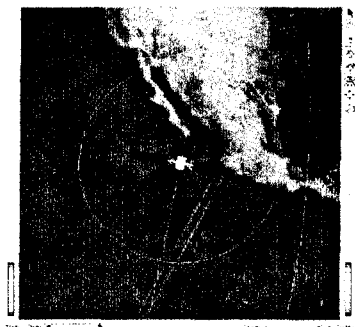


图5 三维可视化视场拉近后的卫星视场

4 结束语

文中空天三维可视化仿真中笔者采用了CPU为P43.0GHz的台式个人计算机,软件平台为Visual Studio.NET 2003,通过Open Inventor三维工具开发包实现了空天目标的模型构建,取得了优于OpenGL的建模效率。文中三维模型运用于实际空天三维可视化仿真与数据处理项目中,取得了良好的效果,在卫星轨

道、地形数据的驱动下模拟出了观测卫星三维运行的空天场景,画面连续,可切换观测视野,可多视场观察。

参考文献:

- [1] 刘 维,韩 潮.基于Open Inventor的航天可视化系统[J].计算机仿真,2006,23(11):23-26.
- [2] 孙家广.计算机图形学[M].北京:清华大学出版社,1994.
- [3] 阎锋欣,侯增选,张定华,等.Open Inventor程序设计从入门到精通[M].北京:清华大学出版社,2007.
- [4] Hao Robin. Open Inventor 简介[EB/OL]. [4]2007. <http://blog.csdn.net/robinhao/>.
- [5] 郑国芹.三维水下虚拟仿真系统的设计与研究[D].哈尔滨:哈尔滨工程大学,2006.
- [6] Wernecke J. Open Inventor Architecture Group,Open Inventor Mentor[M]. [s. l.]: Addison - Wesley Professional, 1994.
- [7] 郝 伟. The Inventor Mentor - 第三章 节点与组[EB/OL]. 2007. <http://blog.csdn.net/robinhao/>.
- [8] Wernecke J. The Inventor Toolmaker - Extending Open Inventor, Release 2[M]. [s. l.]: Addison - Wesley Professional, 1994.
- [9] 李亚臣,蒋红柳,熊海林,等.视景仿真中三维地球的建模[J].计算机工程,2007,33(12):225-227.
- [10] 孙洪军,杜道生,李争航.关于地球形状的三维可视化研究[J].武汉测绘科技大学学报,2000,25(2):158-162.

(上接第4页)

度并不是越高越好,因为过高的维度(例如1024维)会引入大量冗余词,降低正负样本间的区分度,反而降低聚类算法的性能,如图4,图5所示。

(4)由于财经类相对教育类拥有更多能够精确描述类别的特征词汇,故财经类的聚类效果优于教育类。

由此可见,正样本包含类别特征词汇越多,聚类效果越好。

5 结束语

鉴于互联网媒体发布信息信息量大、时效性强的特点,文中将基于密度的聚类思想引入CFK-Means算法,提出了全新的DCFK聚类算法,克服了单纯采用CFK-Means算法依赖初始聚类数和初始聚类中心点的缺陷,同时回避基于密度聚类算法速度慢的缺点。与此同时,文中基于DCFK算法构造大规模中文信息聚类模型,并且通过系列实验验证本聚类模型在互联网媒体信息主动热点发现领域的有效性和实用性。

当然基于DCFK算法的大规模中文信息聚类模型还可融入更进一步的自然语言理解技术,对中文词汇

的词性、词义进行分析,例如给动词和名词赋予高权重,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] Jain A K, Farrokhnia F. Unsupervised texture segmentation using Gabor filters[J]. Pattern Recognition, 1991, 24(13): 1167-1186.
- [2] 曾依灵,许洪波.网络热点信息发现研究[J].通信学报,2007,28(12):141-146.
- [3] Han Jiawei, Kamber M. Data Mining Concepts and Techniques[M]. 范 明,孟小峰,等译.北京:机械工业出版社,2006.
- [4] 唐春生,金以慧.一种大规模的递增聚类算法及其在文档聚类中的应用[J].计算机工程与应用,2002,38(11):187-195.
- [5] 唐春生,金以慧.基于聚类特性的大规模文本聚类算法研究[J].计算机科学,2002,29(9):13-15.
- [6] Jain A K, Murty M N, Flynn P J. Data Clustering: A Review[J]. ACM Computing Surveys, 1999, 31(3): 264-323.