第35卷 第19期 计 算 机 工 程 2009年10月
Vol.35 No.19 Computer Engineering October 2009
·软件技术与数据库· 文章编号：1000—3428(2009)19—0036—03 文献标识码：A 中图分类号：TP311.52

基于文本挖掘的话题发现技术

高 妮[1]，周明全[2]，耿国华[1]，王学松[2]，贺毅岳[1]
(1. 西北大学信息学院计算机科学系，西安 710069；2. 北京师范大学信息科学与技术学院，北京 100875)

摘 要：在分析灾害新闻特点的基础上，提出一种基于文本挖掘的话题发现技术，采用基于平均分组的层次聚类算法，对灾害新闻资料进行组织，从而生成新闻专题，为用户提供个性化服务，并形成专题检测系统，同时介绍基于时间和地点权值向量的相似度计算模型以及基于时间的动态阈值模型。实验结果表明，该算法能够获得较好的性能。

关键词：话题发现与跟踪；层次聚类；文本挖掘；动态阈值

Topic Detection Technique Based on Text Mining

GAO Ni[1], ZHOU Ming-quan[2], GENG Guo-hua[1], WANG Xue-song[2], HE Yi-yue[1]
(1. Dept. of Computer Science, Institute of Information, Northwest University, Xi'an 710069;
2. School of Information Science & Technology, Beijing Normal University, Beijing 100875)

【Abstract】 On basis of analyzing the character of disasters news, a topic detection technique based on text mining is proposed, which uses Group Average Clustering(GAC) algorithm to organize the disasters news materials, generate the news special topics, provide the personality service, and shape the whole system. The similarity computing model based on both weight vectors of time and place and dynamic threshold model based on time are introduced. Experimental results show this algorithm can obtain better performance.

【Key words】 topic detection and tracking; hierarchical clustering; text mining; dynamic threshold

1 概述

2006年，教育部、财政部启动“综合风险防范关键技术研究与示范”项目，其中一项子项目为“综合风险数据搜索和网络信息服务技术”，主要针对综合风险相关信息的突发事件、热点新闻以专题的形式服务于广大网络用户。话题发现与跟踪(Topic Detection and Trac-king, TDT)[1]技术可以将关于事件分散的信息有效地汇集并组织起来。TDT所关注的是对热点新闻、突发事件话题进行组织。文本挖掘也称智能文本分析。文本数据挖掘或文本知识发现是指从非结构化的文本集合中提取有趣的、非平凡的信息和知识。聚类在文本挖掘中是种有用的技术，可以发现有趣的数据分布和基本数据的模式，并在不依靠任何背景知识下可以找到有趣的结构或集群。文本挖掘可以采用聚类、分类算法。

目前，聚类技术[2]大致分为2类：分组化和层次化。层次聚类算法是种理想的交互式的可视化和浏览工具，可以识别出同一粒度层次的话题，具有一致性、可预见性。本文重点是把文档聚类作为一个文件浏览方法。对于话题检测技术，采用基于平均分组的层次聚类(Group Average Clustering, GAC)算法。GAC是针对回溯检测(Retrospective Event Detection, RED)的一种较好算法。

本文提出一种GAC算法，用多层聚类产生层次。每个层次再使用分组聚类操作类簇。挖掘文本内容是从上层开始，用户可以浏览并选择感兴趣的话题。因此，用户在一个全局的视野内浏览各个时间段收集信息，新的文本随着时间的改变可以包含在文档中。

2 专题检测系统的整体架构

本文网络信息服务新闻专题检测系统的整体架构如图1所示。

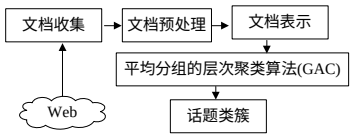

图1 专题检测系统的整体架构

从收集文档开始，该系统再对文档进行预处理，采用改进的文档表示法，根据语义聚集出恰当的话题，然后用基于平均分组的层次聚类算法获得话题结构。

定义1 新闻事件(news event)指发生在某时某地的一件特别的事情。可以用许多属性来描述一个事件，包括事件名称(title)、新闻时间(time)、事件人物(human Name)、事件地点(place name)。

定义2 文档空间(doc)是由一组词和权重对组成的，表示为：$\boldsymbol{doc}=\{(term_1,weight_1),(term_2,weight_2),\cdots,(term_i,weight_i)\}$。文档向量的大小表示为$\|\boldsymbol{doc}\|$。

定义3 事件时间($time_{NE}$)为一时间区间$[time_a \sim time_b]$。其中，$time_a$为最早报道该事件的时间；$time_b$为报道该事件最近的文档时间。

2.1 文档预处理和表示

2.1.1 文档预处理过程

在文件中，代表同一命名实体概念C_i的词通常不止一个，而是个词的集合$C_i=\{t_{i1},t_{i2},\cdots,t_{ik}\}$。例如，代表“中国国

基金项目：国家科技支撑计划基金资助项目(2006BAD20B02)
作者简介：高 妮(1982－)，女，硕士研究生，主研方向：网络信息处理，数据挖掘；周明全、耿国华，教授、博士生导师；王学松，博士研究生；贺毅岳，博士研究生
收稿日期：2009-04-30 **E-mail：**qiuyetingfeng@126.com

家地震局”这个概念的命名实体可能有“国家地震局”、“中国地震局”、“国家地震监测局”等，这种现象在地区及简称方面尤为常见。对代表相同概念的命名实体建立同义库，库中主要收录国家、地区、城市、团体机构等的别称和简称；同时还建立命名实体停用后缀表，后缀如“省”、“市”、“县”、“村”、“地区”等，出现在表中的命名实体后缀被过滤掉。平台中新闻专题检测系统，用爬虫程序在指定的门户网站提取 305 243 篇中文文本的语料库进行训练，命名实体词出现 12 357 677 次，利用同义命名同义起来的就有 1 359 344 个，占命名实体词的 11%；被还原为命名实体的普通词有 865 037 个，占命名实体词的 7%。

2.1.2 改进的文档表示法

从记者的角度看，新闻故事描述一个话题通常会在指定的以下信息：什么时候发生，谁涉及，在哪里发生的以及它如何发生。这些属性是非常有用的描述和总结话题，因此，本专题检测系统为每条新闻的 4 个问题各自分配一个语义组，提出一个专业的表示法：建立 4 个特征向量来选取特征。如 $tags_{term}$={*Human Name*,*Place Name*,*Time*,*Content*}，相应地，每组标记可以用一组权重向量表示 $\boldsymbol{T}^i=(w_1^i,w_2^i,\cdots,w_n^i)$ $(1\le i\le 4)$，这组标记的权重定义为

$$w=\{w_\Lambda \mid \Lambda\in tags_{term}\}$$

通常词对文档的支持度 ($weight_{t_i,d}$) 可以通过词频(TF)和倒排文档频率(IDE)计算[3]。

$$weight'_{t_i,d}=\sum_{\Lambda} tf_{t_i,\Lambda}\times w_\Lambda \tag{1}$$

其中，$tf_{t_i,\Lambda}$ 是被 Λ 标记的词 t_i 在文档中的频率；w_Λ 是 Λ 的权重。对式(1)进行归一化处理：

$$weight_{t_i,d}=\frac{weight'_{t_i,d}}{\max\limits_{t_i ind}\{weight_{t_i,d}\}} \tag{2}$$

出版日期的向量可以从文档中提取。在前期工作中，使用带权重时间的向量代替出版日期。这些时间实体既是日期也是日期间隔，可以通过由 Llido 等人提出的算法进行提取。地点加权重向量可表示为 $(TF_{P_1^i},TF_{P_2^i},\cdots,TF_{P_j^i})$，其中，$TF_{P_k^i}$ 是地点 P_K 在文档 d^i 中的绝对频率。通过 NIMA(the National Imagery and Mapping Agency files)建立的词库，地名可以自动从文档中提取，每个地名用一个带不同地理区域的编码通道包括国家、行政区域等。

2.2 文档相似度计算方法

自动文档聚类是基于对相似度计算和聚类形成的标准，最广泛应用的文档相似度计算方法[4]是余弦计算。比较 2 个文档向量 $\boldsymbol{d}^i$ 和 $\boldsymbol{d}^j$，使用传统的余弦计算：

$$S_T(\boldsymbol{d}^i,\boldsymbol{d}^j)=\cos(\boldsymbol{T}^i,\boldsymbol{T}^j)=\frac{\sum_{k=1}^n w_k^i w_k^j}{\sqrt{\sum_{k=1}^n (w_k^i)^2}\sqrt{\sum_{k=1}^n (w_k^j)^2}}$$

根据 2 个文档的时间向量，提出以下定义：

$$D(\boldsymbol{d}^i,\boldsymbol{d}^j)=\min_{f^i\in FR^i,f^i\in FR^j}\{d(f^i,f^j)\}$$

其中，$d(f^i,f^j)$ 是 2 个文档出版日期相差的天数；FR^i 是文档 $\boldsymbol{d}^i$ 的出版日期 f^i 的集合。

比较 2 个文档的地点向量，提出遵循布尔规则的定义：

$$S_p(\boldsymbol{d}^i,\boldsymbol{d}^j)=\begin{cases}1 & \text{如果}\exists p_q^i,p_l^j\text{, 例如有共同的前缀或未提到地点}\\ 0 & \text{其他}\end{cases}$$

例如：“科索沃”用 5G.02 表示，“南斯拉夫”用 5G 表示。因为它们有共同的前缀，所以相似。这里认为 2 个文档的地点向量相似必须是 2 个文档至少有 1 个地点属于相同的城市。但若文档中没有提到地点，也认为是相似的。最后，整体的相似计算可被定义为

$$S(\boldsymbol{d}^i,\boldsymbol{d}^j)=\begin{cases}S_T(\boldsymbol{d}^i,\boldsymbol{d}^j) & \text{如果}S_P(\boldsymbol{d}^i,\boldsymbol{d}^j)=1\wedge D(\boldsymbol{d}^i,\boldsymbol{d}^j)\le\beta_{time}\\ 0 & \text{其他}\end{cases}$$

其中，β_{time} 是 2 个文档相差的最大天数，它决定 2 个文章是不是同一个主题。在 2 个文档报道同一个话题时，这种相似计算应该有很高的语义相似性：相近的时间和巧合的地点。时间和地点向量被用来做整体相似计算方法的过滤器，可以减少聚类算法的复杂性。

3 文档聚类算法

传统的层次聚类算法目的是建立一个等级文档的层次。然而在实际应用中，各级产生的层次文档并不能达到所需的抽象水平。所以，在层次聚类算法中，每个层次构成文档的抽象类簇。

3.1 类簇的表示

一个类簇的 c 用 $\bar{c}$ 标注，每组标记人物、时间、地点、内容可由元组 (T^c,F^c,P^c) 表示。其中，$T^{\bar{c}}=(T_1^{\bar{c}},T_2^{\bar{c}},\cdots,T_n^{\bar{c}})$，在类簇文档中 $T_j^{\bar{c}}$ 是权重向量 $\boldsymbol{t}_j$ 的平均权重。$F^{\bar{c}}=(F_{f_1}^{\bar{c}},F_{f_2}^{\bar{c}},\cdots,F_{f_s}^{\bar{c}})$，在类簇中 $F_{f_j}^{\bar{c}}$ 是时间向量 $\boldsymbol{f}_j$ 的绝对频率，即这个群集的文档个数包含这个日期，且 s 是描述这个话题类簇发生的总天数。$P^{\bar{c}}=(P_{P_1}^{\bar{c}},P_{P_2}^{\bar{c}},\cdots,P_{P_l}^{\bar{c}})$，在类簇中 $P_{P_j}^{\bar{c}}$ 是地点向量 $\boldsymbol{p}_j$ 的绝对频率，l 是这个群集文档所提到地点的总个数。

3.2 GAC 算法[5]

GAC 算法是针对回溯检测(RED)的一种较好的算法，采用自底向上的贪心算法以及分而治之的策略，能最大化话题类簇中的各新闻报道之间的平均相似度。

3.2.1 GAC 算法的要求

采用平均分组的层次聚类算法必须符合以下要求：

(1)它必须使用相似度涉及时空向量的文档和表示方法。

(2)它的输入必须按照时间排好序的新闻文档集合。

(3)它的输出为层次式的话题类簇结构。

3.2.2 GAC 算法的过程

GAC 算法流程如图 2 所示。

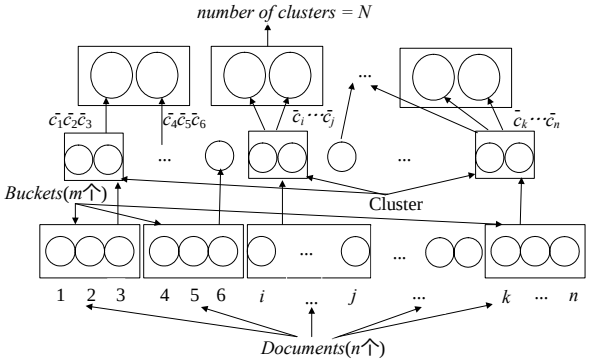

图 2 GAC 算法流程

对图 2 的说明如下：

(1)将文档集合中的每篇文档作为一个单独的话题类簇，初始划分所有单篇文档组成的话题类簇。

(2)将当前话题类簇集合中的话题类簇按顺序连续并且不重叠地划分到 m 个桶中。

(3)对每个桶分别进行聚类。重复地合并桶中的 2 个最相似的低层次话题类簇，形成一个高层次的话题类簇。直到桶中类簇数量减少的比例达到预设的 p 为止，或者任何 2 个类簇之间的相似度值均低于一个预定义阈值 s 为止。

(4)在保持各话题类簇时间顺序的前提下去除桶的边界，也即汇集所有桶中的话题类簇。此时对文档集合的划分即为当前类簇集合；重复(2)~(4)，直到最顶层的话题类簇数目达到了一个预定的数值为止。

(5)定期地将每个顶层类簇中的所有新闻文档按照前 4 步重新聚类。

GAC 算法的时间复杂度为 $O(mn)$，其中，n 为新闻文档集合中的文档数量；m 为桶的大小，且 $m \le n$。该算法不仅效率高，而且考虑新闻文档的时间顺序，提高了话题类簇的质量，通过调整该算法中用到的参数可以改善检测结果。

3.2.3 GAC 算法阈值的判断

在 GAC 算法流程的说明(3)中，一个预定义阈值 s 是动态改变的。本文检测系统采用基于时间的阈值模型。在时间上距离某个话题越远的新闻报道越难加入该话题。如果文档和事件之间的相似度小于 s，那么该新闻报道是新事件。阈值 s 的选择很关键。直接指定一个常数作为阈值是不科学的，它应该与具体的文档和事件相关，而 $S(\boldsymbol{d}^i,\boldsymbol{d}^j)$ 体现了文档和事件之间的相关长度，这里，$threshold(\boldsymbol{d}^i,\boldsymbol{d}^j) = 0.4 + \alpha[S(\boldsymbol{d}^i,\boldsymbol{d}^j) - 0.4] + \beta(date^j - date^i)$，其中，$S(\boldsymbol{d}^i,\boldsymbol{d}^j)$ 为当前新闻报道与话题类簇之间的相似度；$(date^j - date^i)$ 为新闻报道的到达时间与话题类簇创建之间间隔的天数；α，β 为调整因子；α 表示基于内容相似度的影响，取值在 $[1,2]$ 之间；β 表示时间距离的影响，取值在 $[0,0.1]$ 之间；0.4 为 Inquery 系统中的经验值。

4 实验结果

平台中新闻专题检测系统，用爬虫程序在指定的门户网站提取 305 243 篇中文文本的语料库进行训练。涉及具有代表性的事件有 3 个：2008 南方雪灾，手足口病，2008 汶川地震等事件，如表 1 所示。

表 1 实验语料相关信息统计

事件 *ID*	事件名	起始日期	终止日期	文档数
*NE*1	2008 南方雪灾	2008-01-10	2008-02-02	25 797
*NE*2	手足口病	2008-03-07	2008-04-26	3 998
*NE*3	2008 汶川地震	2008-05-12	未知	52 183
…	…	…	…	…

在 TDT 中，使用失报(misses)率和错报(false alarms)率评价话题发现的效率，而本文使用 2 种传统的评价标准：F1-measure 和检测费用。F1-measure 被广泛地应用于话题检测系统中，它由召回(recall)率以及准确(precision)率进行综合考虑。

召回率：$R = \frac{a}{a+c}$；正确率：$P = \frac{a}{a+b}$；F1-measure：$\frac{2PR}{P+R}$；失报率：$\frac{c}{a+c}$；错报率：$\frac{b}{b+d}$。

其中，a 为话题发现的相关新闻文档数；b 为话题发现的不相关新闻文档数；c 为未发现的相关新闻文档数；d 为未发现的不相关新闻文档数。

对原始数据进行 3 次实验：(1)在计算文档和话题之间相似度时未考虑时间，地点权重向量(α =1.6 和 β =0)；(2)加入时间权重(α =1.7 和 β =0.002)；(3)加入时间和地点权重向量(α =1.8 和 β =0.005)。如图 3 所示。

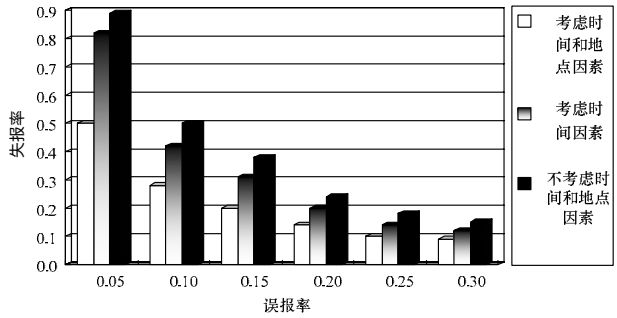

图 3 失报率和错报率在话题中的效率值

从图 3 可以看出，时间和地点权重向量的相似度计算对话题发现的失报率和误报率都有所降低。例如事件“2008 汶川地震”在误报率为 20%的情况下，考虑到时间和地点因素的话题发现失报率为 14%，考虑到时间因素的话题发现失报率为 20%，而没有考虑时间和地点因素的话题发现失报率为 24%。

表 2 显示了时间、地点因素对话题发现效率所造成的影响。基于时间、地点权值向量的相似度明显比基于内容相似度的话题发现效率有了很大程度的提高。实验证明在计算文档和事件之间的相似度时，考虑时间和地点因素，可以得到令人满意的结果。

表 2 F1-measure 在话题中的效率值

事件 *ID*	事件名	F1(不考虑时间和地点因素)	F1(考虑时间因素)	F1(考虑时间和地点因素)	文档数
*NE*1	2008 南方雪灾	0.679 2	0.812 2	0.862 9	25 797
*NE*2	手足口病	0.588 4	0.693 8	0.725 6	3 998
*NE*3	2008 汶川地震	0.639 4	0.766 9	0.809 7	52 183
…	…	…	…	…	…

5 结束语

本文结合“综合风险数据搜索和网络信息服务技术”项目中的新闻专题检测系统的需求，提出基于文本挖掘的话题发现技术，包括基于时间和地点权值向量的相似度计算模型、基于时间的动态阈值模型和采用平均分组的层次聚类算法实现话题发现，实验结果表明，GAC 算法提高了新闻事件的发现性能，且是有效的。

参考文献

[1] Margaret C. Intelligence Information Retrieval[C]//Proc. of the 7th Int'l Conf. on Topic Detection and Tracking. Gaithersbury, USA: [s. n.], 2004.

[2] Broder A Z, Glassman S C, Manasse M S, et al. Syntactic Clustering of the Web[C]//Proceedings of the 6th International Web Wide World Conference. [S. l.]: ACM Press, 1997.

[3] 李晓黎, 刘继敏, 史忠植. 基于向量机和无监督聚类相结合的中文网页分类器[J]. 计算机学报, 2001, 24(1): 62-68.

[4] 贾自艳, 何 清, 张海俊, 等. 一种基于动态进化模型的事件探测和追踪算法[J]. 计算机研究与发展, 2004, 41(7): 33-35.

[5] 骆卫华, 于满泉, 许洪波, 等. 基于多策略优化的分治多层聚类算法的话题发现研究[J]. 中文信息学报, 2006, 20(1): 29-36.

编辑 陈 文