仲兆满,李存华,戴红伟,等.融合内容与时间特征的中文新闻子话题聚类[J].计算机科学与探索,2013,7(4):368-377.

ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2013/07(04)-0368-09
doi: 10.3778/j.issn.1673-9418.1205024

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-51616056

融合内容与时间特征的中文新闻子话题聚类*

仲兆满[1+],李存华[1],戴红伟[1],刘宗田[2]
1. 淮海工学院 计算机工程学院,江苏 连云港 222005
2. 上海大学 计算机工程与科学学院,上海 200072

Clustering Chinese News Subtopic Integrating Content and Time Features*

ZHONG Zhaoman[1+], LI Cunhua[1], DAI Hongwei[1], LIU Zongtian[2]
1. School of Computer Engineering, Huaihai Institute of Technology, Lianyungang, Jiangsu 222005, China
2. School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China
+ Corresponding author: E-mail: zhongzhaoman@163.com

ZHONG Zhaoman, LI Cunhua, DAI Hongwei, et al. Clustering Chinese news subtopic integrating content and time features. Journal of Frontiers of Computer Science and Technology, 2013, 7(4): 368-376.

Abstract: Subtopic is the division for the topic, and it is a new research direction compared with the topic. Subtopic clustering is the base for the analysis of topic evolution relations. This paper proposes a new method of clustering Chinese news subtopic integrating content and time features. It focuses on the analysis of subtopic content feature in text, and studies the computation of subtopic word weights and the dimension reduction of subtopic words. Five topics including 18 subtopics are used to conduct the experiment. Experimental results show that the performance of the proposed method is better than the existing subtopic identification methods.
Key words: topic evolution; subtopic clustering; content features; time features

摘 要:子话题是对话题的再次划分,是比话题粒度更细的新兴研究方向,子话题的聚类是话题内部演化关系分析的基础。提出了融合内容特征和时间特征的中文新闻子话题聚类方法,重点分析了子话题内容特征的表现规律,研究了子话题特征词的权重计算和降维方法。选取5个话题的18个子话题进行了实验,结果表明,所提方法的性能与已有的子话题聚类方法相比有显著提高。
关键词:话题演化;子话题聚类;内容特征;时间特征
文献标志码:A **中图分类号:**TP311.5

* The National Natural Science Foundation of China under Grant No. 60975033 (国家自然科学基金); the Natural Science Foundation of Lianyungang of China under Grant No. CG1121 (连云港市自然科学基金).
Received 2012-05, Accepted 2012-07.

1 引言

由于互联网信息的爆炸性、异构性、分布性，与一个话题（topic）相关的信息往往分散在很多不同的地方。如何把握话题内容的发展趋势和话题的迁移过程，了解事件在不同阶段的热点关注问题，是很难做到的。目前的各种信息检索、分类、监测和提取技术都是围绕这个目的展开的[1-3]。

话题检测与跟踪（topic detection and tracking，TDT）是一种以话题为主线，对信息进行组织汇总的技术，旨在依据事件对新闻数据流进行组织和利用，也是为了应对信息过载问题而提出的一项应用研究。然而，话题跟踪把话题中的所有报道简单地看成一个文本集合，TDT的研究结果只是将和话题相关的报道聚集到了一起，没有考虑话题内部的演化特征，无法体现出话题中各事件的来龙去脉。

为了更好地帮助用户了解话题的进展情况，需要将和一个话题相关的所有报道按照它们所关注的内容的不同再次进行划分，即子话题。话题内部子话题之间的演化关系分析超出了TDT研究的范畴，是一个新的研究方向，正在引起研究者们的关注。子话题研究工作的进展将改变目前搜索引擎、信息监测、信息分类等系统单纯地返回报道集合，缺乏报道集合内事件演化趋势分析的缺陷。

子话题的聚类是分析话题中子话题的演化趋势的基础工作。本文研究的出发点是：假设一个话题已经获取了相关的报道，需要聚类得到一个话题中的若干个子话题。这一工作主要涉及子话题特征选取，子话题聚类等研究内容。

本文组织结构如下：第2章介绍了相关工作及存在的问题；第3章阐述了话题、子话题和事件的内涵；第4章分析了子话题的内容和时间特征；第5章提出了融合内容与时间特征的中文新闻子话题的聚类方法，并进行了实验分析；最后对全文进行了总结。

2 相关工作

2.1 话题表示模型

目前TDT研究中常用的报道和话题表示模型有语言模型和向量空间模型，子话题也大多借鉴了这两种表示模型。假设报道中出现的词 t 互不相关，则某篇报道特征与话题特征的相关概率为：

$$P(T|d)=\frac{P(T)\times P(d|T)}{P(d)}\approx P(T)\prod_{n}\frac{P(t|T)}{P(t)} \tag{1}$$

其中，$P(T)$ 是一篇报道和话题 T 相关的先验概率；$P(t|T)$ 表示词 t 在某话题 T 中的生成概率；$P(t)$ 是词 t 在整个语料库中的分布概率。话题语言模型一般很稀疏，需要解决未见词的0概率问题，通常采用线性插值 $\lambda P(t|T)+(1-\lambda)P(t)$ 把背景语言模型加入进去，为了减少跟踪代价，一些跟踪系统中系数 λ 取值为0.25[4]。

向量空间模型是目前TDT研究中最常用的报道表示模型，即将话题形式化为多维空间中的一个点，以向量的形式给出。为了更加准确地表示报道内容，有些研究者们使用多个向量空间模型来表示一篇报道。比如Lam等人将报道表示成命名实体向量和内容向量两个向量[5]；而文献[6]则将报道表示成四个向量：地点向量、时间向量、名字向量和内容向量。

向量空间模型对项的权重评价、相似度的计算没有进行统一的规定，只是提供一个理论框架，可以使用不同的权重评价函数和相似度计算方法，因此该模型有广泛的适应性。当前使用最多的权值评价方法是 $TF\times IDF$ 。词语频次（term frequency，TF）指词语在文档中出现的次数，词语倒排文档频次（inverse document frequency，IDF）是词语在文档集合中分布情况的一种量化，词 t 的 IDF 常用的计算方法是 $\mathrm{lb}(N/n_t+0.01)$，其中 N 为文档集合中文档的数目，n_t 为出现词 t 的文章数。根据 TF 和 IDF 两个因素，得到词 t 的权重计算方法如式（2）所示：

$$w_t=TF_t\times\mathrm{lb}(N/n_t+0.01) \tag{2}$$

其中，w_t 为词 t 的权重；TF_t 为词 t 的频次。

通常要对向量进行归一化，归一化方法如式（3）所示：

$$w_t=\frac{TF_t\times\mathrm{lb}(N/n_t+0.01)}{\sqrt{\sum_{k=1}^{m}(TF_k\times\mathrm{lb}(N/n_k+0.01))^2}} \tag{3}$$

2.2 子话题聚类

要想对子话题之间的演化关系进行分析，首先

必须聚类得到话题中包含的子话题。IBM公司围绕话题识别开发了一个相对比较成功的系统。该系统采用了两层聚类策略,使用对称的Okapi公式来比较两篇报道的相似性。该系统首先将报道暂时归入不同的小话题簇,然后在有限的延迟时间后再将其归入最终的话题簇。

王巍根据搜索引擎返回的有关某个话题的结果进行子话题划分,提出了两种子话题聚类方法,分别是基于关键词的划分方法和基于时间信息的划分方法[7]。基于关键词的划分方法中,首先计算关键词的权重,然后根据关键词进行分类,但是并没有分析子话题的内容特征。在基于时间信息的划分方法中,简单地将同一时间点的子话题片段进行合并,并没有考虑同一时间点可能出现多个子话题的情况。

张阔等人提出了基于关键词元的话题内事件检测方法,只是检测事件,并未形成子话题,而且仅在英语语料上进行了测试[8]。

文献[6]针对话题分析第一次提出了动态演化的概念,并提出使用本体,包括一些人名、地名、时间等词汇,来测量事件之间的相似度。但是该文并没有对演化分析涉及的子话题进行更加细致的定义,也没有提供任何实验结果。

文献[9]提出了基于子话题分治匹配的新事件检测方法,将话题和报道划分为不同子话题,根据相关子话题的比例关系和分布关系建立新话题识别模型。

已有的子话题聚类的研究存在的主要问题如下:

(1)在一定的时间段内,会存在多个子话题,基于时间信息的子话题聚类方法克服不了这一问题。比如"地震"话题出现后,在较短的时间内会引起"救援"、"人员伤亡"、"交通中断"、"捐款捐物"等子话题。

(2)已有的基于关键词的子话题聚类研究,没能分析一个话题的子话题特征词的分布特性,进而使用合适的子话题特征选择策略。

(3)已有研究工作大多以英文的话题报道为研究对象,对中文领域的研究关注得不够。

3 话题、子话题及事件的内涵

话题是话题跟踪研究中的一个基本概念,它的含义与语言学上使用的概念不同。在最初的研究阶段,话题和事件含义相同[10]。一个话题指由某些原因、条件引起,发生在特定时间、地点,有一定的参与者或涉及者,并可能伴随某些必然结果的一个事件。比如"2008年5月12日汶川发生了地震"。目前使用的话题概念比较宽泛,它包括一个核心事件或活动,以及所有与之直接相关的事件或活动。也可以说话题由一个种子子话题和其他相关子话题构成。

话题的子话题划分有利于分析话题的内容构成,包括某个时期关注的内容和不同时期不同内容之间的关系,有利于建立话题的演化模型,研究话题的发展趋势,更加清晰地了解互联网话题信息的构成情况。

根据话题的定义,一篇报道只要论述的事件或活动与一个话题的种子事件有着直接的联系,这篇报道就与该话题相关。比如"灾后重建"与"捐款捐物"的报道都认为与"汶川地震"事件直接相关,因此可以作为该话题的一个组成部分。

一个话题的子话题构成随着时间的推移会发生变化。例如话题"汶川地震"经历了从"灾区救援"、"捐款捐物"到"灾后重建"的演化过程。可见,话题的分析不能只停留在静态话题内容的提取分类上,还要动态分析话题中子话题的演化过程,找到子话题间的相关特性。利用话题内部的演化模型,可以对一些突发事件进行前期预测及控制,降低一些不良话题的爆发概率。

事件指在某个特定时间和环境下发生的,由若干角色参与,表现出若干动作特征的一件事情[11]。形式上,事件可表示为 e ,定义为一个六元组 $e=<A,O,T,V,P,L>$ 。其中,事件六元组中的元素称为事件要素,分别表示动作(A)、对象(O)、时间(T)、环境(V)、断言(P)和语言表现(L)。

话题、子话题及事件之间的关系如图1所示。

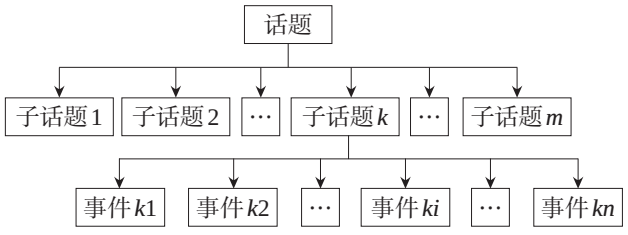


Fig.1 Relations of topic, subtopic and event

图1 话题、子话题及事件之间的关系

图1描述了话题、子话题和事件之间的关系，它们的关系形式化定义为：$T=\{ST_1, ST_2, \cdots, ST_m\}$，$ST_k=\{E_{k1}, E_{k2}, \cdots, E_{kn}\}(1\leqslant k\leqslant m)$。其中，$T$ 指话题，由 m 个子话题组成；ST_k 指某个子话题，由 n 个事件组成。

新闻报道对话题的阐述总是围绕若干事件展开的。如果一篇新闻报道 d 介绍的是同一个子话题 ST_k 的若干个事件，则 d 属于子话题 ST_k；如果一篇新闻报道 d 介绍的是不同子话题的若干个事件，则 d 可以属于多个不同的子话题，即新闻报道和子话题间是多对多的关系。

4 子话题的内容与时间特征

4.1 子话题的内容特征

子话题在内容上具有如下特征：

(1)种子事件贯穿于各个子话题，而且出现的频次比较高。种子事件对区分不同的话题比较有效，而对区分一个话题的不同子话题作用不大，相反会增加报道之间的相似度，影响子话题聚类的效果。比如对话题“汶川地震”而言，种子事件是“汶川地震”，在“人员伤亡”、“捐款捐物”、“反贪调查”等子话题中都频繁出现。在对子话题进行特征提取时，应该删除种子事件的干扰，这包括种子事件的事件触发词、地点、时间及参与者等要素。比如话题“汶川地震”，种子事件是“5月12日汶川地震”，应该删除地点要素“汶川”、时间要素“5月12日”及“地震”事件触发词要素。

(2)子话题是通过各类事件描述的，而事件通常关联了事件触发词、时间、地点、参与者等要素，对子话题而言事件触发词相比较其他要素具有更重要的区分能力。比如子话题“捐款捐物”，关注的是与“捐款捐物”相关的事件，并不关注哪个组织或个人捐款捐物。除种子事件外，其他事件触发词代表了一类事件，并不特指某个事件，只有在时间、地点、参与者等要素明确的情况下才联合表征特定事件。对事件触发词要素应赋予更高的权重，降低时间、地点、参与者等特定要素的权重。这与常用的话题聚类技术也是完全不同的，常用的话题聚类技术对事件触发词、人物、时间和地点等要素都赋予了较高的权重[12-14]。

(3)新闻报道的标题、开始的几个句子往往就能交代所要表述的子话题，就能代表报道的主题。特征词出现在不同的位置，其重要程度有所不同，对于出现在新闻报道的标题、开始的几个句子的特征词应该赋予更高的权重。

4.2 子话题的时间特征

时间可以反映话题发展的趋势。比如话题的种子事件发生的时间最早，并长期驻留于相关的话题报道流中。与此相对地，话题的新颖事件往往发生的时间较晚，但论述新颖事件的报道会在短期内有爆发式的增益。同时，时间也是区分不同事件的重要属性。其依据在于事件定义为特定时间发生的事情，换言之，不同时间发生的事件不是同一事件。新闻信息的这一特性有助于子话题的聚类。

互联网新闻报道有很强的时间时序性，多是以时间(大部分最小单位为“日”)为基本叙述单元，而且时间的表述方式结构相对简单。新闻报道中可以利用的时间分为两类：报道的发表时间和报道中包含的事件时间。本文侧重于使用报道的发表时间。报道的发表时间一般位于新闻标题的下方，或者是正文的结束位置，在文本中的表现形式也比较单一。发表时间可以分为绝对时间和相对时间两种。绝对时间指明确的报道发表的具体时间，比如“2012-3-20”、“2012/3/20”、“2012:3:20”等，在报道中的表现格式比较单一。而相对时间指报道的发表时间是相对于系统的当前时间，表现的格式稍微复杂。经统计汇总后，常见的新闻报道发表的相对时间格式如表1所示。

Table 1 Relative published time formats of news

表1 新闻报道发表的相对时间格式

序号	时间格式	例子	序号	时间格式	例子
1	秒前	20秒前	7	今天	今天
2	分钟前	5分钟前	8	天前	2天前
3	半小时前	半小时前	9	秒之前	10秒之前
4	小时前	2小时前	10	分钟之前	30分钟之前
5	昨天	昨天	11	小时之前	1小时之前
6	前天	前天			

5 融合内容与时间特征的中文新闻子话题聚类方法

5.1 方法流程

子话题聚类的任务是对一个话题所有的报道进行再次划分。假设和话题 T 相关的报道的集合为 $D=\{d_1, d_2, \cdots, d_n\}$，则子话题聚类的结果是建立一个子话题集合 $T=\{ST_1, ST_2, \cdots, ST_m\}$。其中每个子话题都非空，并且满足以下关系：$\forall d_i, \exists ST_j \in T, d_i \in ST_j$ $(1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant m)$。

融合内容与时间特征的中文新闻子话题聚类方法的流程如图2所示。

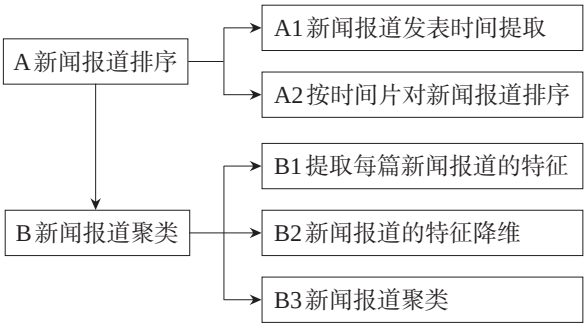


Fig.2 The flow of clustering Chinese news subtopics

图2 中文新闻子话题聚类方法流程

图2所示的方法流程主要包括两个步骤：新闻报道排序(A)和新闻报道聚类(B)。

5.1.1 新闻报道排序

步骤A新闻报道排序包括新闻报道发表时间提取(A1)和按时间片对新闻报道排序(A2)两个子步骤。步骤A1提取新闻的相对发表时间和绝对发表时间，将相对发表时间转化为绝对发表时间，并统一形式化为：$Time=\{$年,月,日$\}$。步骤A2以“天”为单位对一个话题中所有的新闻报道升序排序。假设新闻报道的集合为 $D=\{d_1, d_2, \cdots, d_n\}$，按照时间以“天”为单位对 D 排序后，得到的结果为 $D=\{D_1, D_2, \cdots, D_h\}$，$h$ 为 D 中新闻报道跨越的天数，其中 $D_i=\{d_{i1}, d_{i2}, \cdots, d_{ij}\}$ $(1 \leqslant i \leqslant h)$，$j$ 为第 i 天 D 中新闻报道的篇数。

5.1.2 新闻报道聚类

步骤B包括提取每篇新闻报道的特征(B1)、新闻报道的特征降维(B2)和新闻报道聚类(B3)三个子步骤。步骤B1在提取每篇新闻报道的特征时，使用向量空间模型(vector space model，VSM)表示文本的特征。特征词 t 的权重一方面取决于特征词在新闻报道中出现的次数及位置，该权重记做 w_t^{TF}；另一方面取决于特征词的倒排文档频次，该权重记做 w_t^{IDF}。在权值计算时，出现在新闻报道的标题中的词被认为是第一个等级，正文中的第一个句子(句子以“。”、“!”、“?”为结束标记)是第二个等级，其他的句子依次类推。句子越后出现，句子中的词的权值也越小。w_t^{TF} 的计算方法如式(4)所示：

$$w_t^{\mathrm{TF}}=\sum_{k=1}^{m}\frac{1}{2^{\ln t_k}} \tag{4}$$

其中，m 为特征词 t 在报道中出现次数；t_k 是特征词 t 的第 k 次出现所处的等级，取对数可以降低相邻两个等级间的差距。

w_t^{IDF} 的计算方法如式(5)所示：

$$w_t^{\mathrm{IDF}}=\mathrm{lb}(N/n_t+0.01) \tag{5}$$

其中，N 为报道总数；n_t 为包含特征词 t 的报道数。

最后得到特征词 t 的权重如式(6)所示：

$$w(t)=w_t^{\mathrm{TF}}\times w_t^{\mathrm{IDF}} \tag{6}$$

步骤B2在对每篇新闻报道的特征进行降维时，事件触发词的识别采用文献[15]介绍的方法。该方法将停用动词分为两大类：第一类动词，不作为事件触发词，直接舍弃，常见的是存现动词、能愿动词、判断动词、使令动词、主观感觉、猜想、阐述等动词；第二类动词，不作为事件触发词，但与其一起出现的动词、名动词或者名词可以作为事件触发词，比如“发生火灾”、“开始演讲”、“禁止吸烟”等，“发生”、“开始”和“禁止”不是事件触发词，而名词“火灾”、动词“演讲”和“吸烟”是事件触发词。识别报道中的事件触发词后，增大事件触发词的权重，方法如式(7)所示：

$$w_i=w_i\times \mathrm{lb}\,n \tag{7}$$

其中，w_i 是事件触发词 e_i 的权重；n 是 e_i 在文本中出现的次数。

步骤B3对新闻报道的聚类使用经典的 K-means 聚类算法，按照时间次序依次选取新闻报道进行聚类。本文对 K-means 算法不作介绍。

5.2 实验结果及分析

为了验证本文理论分析及所提方法的有效性，在互联网上围绕近期的热点事件选取了真实的话题及子话题进行了实验，并对结果进行了比较分析。选定的话题及确定的子话题如表2所示，括号中的数字表示收集新闻报道的篇数。

表2所示的实验资料是由本科生人工收集整理的，对子话题的确定和报道到子话题的划分都经过了认真的商讨。

5.2.1 子话题内容特征的实验分析

表3列出了话题“汶川地震”的5个子话题中部分特征词按照权重进行的降序排序结果（每个子话题取前10个特征词）。特征词的权重计算使用 *TF*×*IDF* 方法，仅仅进行了停用词过滤，没有进行特征降维。

如表3所见，种子事件的事件触发词“地震”在5个子话题中都拥有较高的权重，排名比较靠前。种子事件的地点要素“汶川”在4个子话题中（“人员伤亡”、“灾区救援”、“捐款捐物”和“反贪调查”）的排名都比较靠前，地点要素“成都”在2个子话题中（“人员伤亡”、“灾区救援”）的排名比较靠前。种子事件的时间要素“5月12日”在2个子话题中（“人员伤亡”、“灾区救援”）的排名比较靠前。

子话题是通过具体的事件而表述的，对表3而言，5个子话题的事件触发词出现得并不多。子话题“人员伤亡”包含3个事件：“地震”、“伤亡”和“报告”；子话题“灾区救援”包含3个事件：“救援”、“地震”和“救灾”；子话题“捐款捐物”包含3个事件：“捐款”、“地震”和“捐物”；子话题“灾后重建”包含6个事件：“重建”、“地震”、“恢复”、“灾害”、“规划”和“建设”；子话题“反贪调查”包含3个事件：“重建”、“捐款”和“地震”。表3中事件触发词在所有词中的比例为

Table 2 Topics and subtopics for the experiment

表2 实验选用的话题及其子话题

话题	子话题
汶川地震（99篇）	①人员伤亡（22篇）；②捐款捐物（16篇）；③灾区救援（20篇）；④灾后重建（26篇）；⑤反贪调查（15篇）
温州动车事故（83篇）	①人员伤亡（18篇）；②事故原因调查（20篇）；③责任追究（21篇）；④事故赔偿（24篇）
911恐怖袭击（111篇）	①基地组织（25篇）；②本拉登（24篇）；③救援（19篇）；④国际影响（25篇）；⑤人员伤亡（18篇）
三鹿奶粉事件（80篇）	①奶源探访（16篇）；②赔偿（24篇）；③问责（21篇）；④官员复出（19篇）
中日钓鱼岛撞船（83篇）	①撞船（26篇）；②强制起诉（19篇）；③巡航（21篇）；④索赔（17篇）

Table 3 Ranking some keywords for 5 subtopics of “Wenchuan earthquake” topic

表3 “汶川地震”的5个子话题部分特征词的排序

人员伤亡		灾区救援		捐款捐物		灾后重建		反贪调查	
特征词	权重	特征词	权重	特征词	权重	特征词	权重	特征词	权重
地震	3 212	救援	2 834	灾区	2 093	重建	6 120	后	1 794
汶川	1 468	灾区	1 950	捐款	2 055	**地震**	5 832	元	1 500
人员	1 357	**地震**	1 071	元	1 442	后	4 522	钱	1 243
伤亡	1 104	人	901	企业	781	恢复	2 340	重建	1 023
级	810	**汶川**	650	**地震**	780	灾害	1 612	办公室	616
报告	608	队	624	中国	671	规划	1 547	工作	488
震感	570	人员	462	公司	660	建设	1 500	捐款	460
大	448	救灾	312	集团	650	人	1 296	证据	450
成都	435	**5月12日**	294	**汶川**	627	大	1 287	**地震**	432
5月12日	430	**成都**	273	捐物	546	设施	1 170	**汶川**	408

Table 4 F -value of topic ‘earthquake’ using three methods

表4 “地震”话题使用三种聚类方法得到的 F 值

子话题	应聚类文本数	M1			M2			M3		
		V	U	F值/(%)	V	U	F值/(%)	V	U	F值/(%)
人员伤亡	22	23	14	62.2	22	13	59.1	21	15	69.8
灾区救援	16	18	11	64.7	15	12	77.4	15	13	83.9
捐款捐物	26	22	17	70.8	25	18	70.6	25	19	74.5
灾后重建	20	17	14	75.7	19	14	71.8	22	17	81.0
反贪调查	15	19	12	70.6	18	12	72.7	16	12	77.4
平均F值				68.8			70.3			77.3

18/50=36% 。对5个话题的所有子话题按照 $TF\times IDF$ 方法计算权重降序排序后,分别取每个子话题的前10个特征词,得到事件触发词在所有特征词中的比例为 92/250=36.8% 。

对5个话题的所有子话题使用本文方法计算权重,即使用式(6),经过过滤种子事件的要素及增大事件触发词的权重后,分别取每个子话题的前10个特征词,得到事件触发词在所有特征词中的平均比例为 141/250=58.4% 。相比较单纯地使用 $TF\times IDF$ 计算特征词的权重,事件触发词在特征词中的比例提高了 21.6% 。

5.2.2 内容特征词的选取对子话题聚类效果的影响

将每个话题的子话题打乱,按照5.1节所示的流程对话题中的语料进行聚类,得到各个子话题。用三种方法进行了实验比较:方法M1表示使用所有特征词,未进行特征词的特殊加权处理;方法M2表示增大特征词中的事件触发词、地点、时间、参与者四类要素的权重;方法M3表示删除特征词中的种子事件各要素,只增大事件触发词的权重。

对实验得到的数据采用的评估指标是 F 值,F 值的计算方法如式(8)所示:

$$F=\frac{P\times R\times 2}{P+R} \tag{8}$$

其中,P 是准确率;R 是召回率。它们的计算方法分别如式(9)和式(10)所示:

$$P=\frac{U}{V} \tag{9}$$

$$R=\frac{U}{W} \tag{10}$$

其中,U 表示聚类正确的报道数;V 表示实际聚类的报道数;W 表示应该聚类的报道数(即一个类别的标准答案)。对“地震”话题,使用三种聚类方法得到的实验结果如表4所示。

对5个话题,使用三种方法得到的平均 F 值如表5所示。

Table 5 F-value of five topics using three methods

表5 5个话题使用三种聚类方法得到的 F 值 (%)

话题	M1	M2	M3
汶川地震	68.8	70.3	77.3
温州动车事故	67.9	71.4	76.4
911恐怖袭击	66.9	72.0	79.1
三鹿奶粉事件	69.1	69.6	75.9
中日钓鱼岛撞船	68.0	69.2	77.5
平均F值	68.1	70.5	77.2

由表4和表5可见,方法M3和方法M1、M2相比,在 F 值方面有较大的提高。方法M3的平均 F 值为77.2%,较之方法M1提高了9.1个百分点,较之方法M2提高了6.7个百分点。方法M2的平均 F 值为70.5%,较之方法M1仅提高了2.4个百分点。

6 结束语

本文针对中文新闻子话题的聚类展开了研究,重点是子话题在内容特征上与话题的异同比较。在融合新闻报道的内容与时间特征的基础上,提出了一种新颖的中文新闻子话题聚类方法,并选用了5个话题进行了实验分析与比较。实验结果证明了所提方法的有效性。子话题的聚类是开展话题内部研究

的基础工作，对人们了解话题组成结构、演化规律具有重要的意义。为了取得更好的子话题聚类效果，需要在分词、事件各个要素的识别，时间格式的规整，子话题的特征权重计算与降维等方面进行更加深入的研究。

References:

[1] Chang T H, Lee C H. Subtopic segmentation for small corpus using a novel fuzzy model[J]. IEEE Transactions on Fuzzy Systems, 2007, 15(4): 699-709.

[2] Zhang Xiaoyan, Wang Ting. Research of technologies on topic detection and tracking[J]. Journal of Frontiers of Computer Science and Technology, 2009, 3(4): 347-357.

[3] Yang C Y, Shi X D, Wei C P. Discovering event evolution graphs from news corpora[J]. IEEE Transactions on Systems, Man, and Cybernetics: Part A Systems and Humans, 2009, 39(4): 850-863.

[4] Ponte J, Croft W. A language modeling approach to information retrieval[C]//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98), Berkeley, USA, 1998. New York, NY, USA: ACM, 1998: 275-281.

[5] Lam W, Meng H M, Hui K. Multilingual topic detection using a parallel corpus[C]//Proceedings of the Topic Detection and Tracking Workshop, 2000.

[6] Makkonen J. Investigations on event evolution in TDT[C]//Proceedings of the Student Workshop of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Edmonton, Canada, 2003. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003: 43-48.

[7] Wang Wei. Analysis of network topic evolution based on keywords and time[D]. Shanghai: Fudan University, 2009.

[8] Zhang Kuo, Li Juanzi, Wu Gang, et al. Term-committee-based event identification within topics[J]. Journal of Computer Research and Development, 2009, 46(2): 245-252.

[9] Hong Yu, Zhang Yu, Fan Jili, et al. New event detection based on division comparison of subtopic[J]. Chinese Journal of Computers, 2008, 31(4): 687-695.

[10] Fiscus J G, Doddington G R. Topic detection and tracking evaluation overview[M]//Topic Detection and Tracking: the Information Retrieval Series. Norwell, MA, USA: Kluwer Academic Publishers, 2002: 17-31.

[11] Liu Zongtian, Huang Meili, Zhou Wen, et al. Research on event-oriented ontology model[J]. Computer Science, 2009, 36(11): 191-195.

[12] Makkonen J, Ahonen-Myka H, Salmenkivi M. Applying semantic classes in event detection and tracking[C]//Proceedings of the 2002 International Conference on Natural Language Processing (ICON 2002), Mumbai, India, 2002: 175-183.

[13] Zhao Yanyan, Qin Bing, Che Wanxiang, et al. Research on Chinese event extraction[C]//Proceedings of the 3rd National Conference on Information Retrieval and Content Security (NCIRCS '07), Suzhou, 2007: 55-62.

[14] Makkonen J, Ahonen-Myka H, Salmenkivi M. Topic detection and tracking with spatio temporal evidence[C]//Proceedings of the 25th European Conference on Information Retrieval Research (ECIR '03). Berlin, Heidelberg: Springer-Verlag, 2003: 251-265.

[15] Zhong Zhaoman, Zhu Ping, Li Cunhua, et al. Research on event-oriented query expansion based on local analysis[J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(2): 151-159.

附中文参考文献：

[2] 张晓艳, 王挺. 话题发现与追踪技术研究[J]. 计算机科学与探索, 2009, 3(4): 347-357.

[7] 王巍. 基于关键词和时间点的网络话题演化分析[J]. 上海: 复旦大学, 2009.

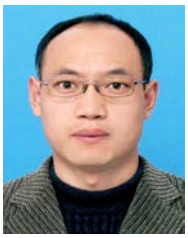
[8] 张阔, 李涓子, 吴刚, 等. 基于关键词元的话题内容事件检测[J]. 计算机研究与发展, 2009, 46(2): 245-252.

[9] 洪宇, 张宇, 范基礼, 等. 基于子话题分治匹配的新事件检测[J]. 计算机学报, 2008, 31(4): 687-695.

[11] 刘宗田, 黄美丽, 周文, 等. 面向事件的本体模型[J]. 计算机科学, 2009, 36(11): 191-195.

[13] 赵妍妍, 秦兵, 车万翔,等. 中文事件抽取技术研究[C]//第三届全国信息检索与内容安全学术会议, 苏州, 2007: 55-62.

[15] 仲兆满, 朱平, 李存华, 等. 一种基于局部分析面向事件的查询扩展方法[J]. 情报学报, 2012, 31(2): 151-159.



ZHONG Zhaoman was born in 1977. He received his Ph.D. degree in computer applications from Shanghai University in 2011. Now he is an associate professor at Huaihai Institute of Technology. His research interests include information retrieval, text information mining and event ontology, etc.

仲兆满(1977—),男,江苏赣榆人,2011年于上海大学计算机应用专业获得博士学位,现为淮海工学院副教授,主要研究领域为信息检索,文本信息挖掘,事件本体等。在国内外期刊和会议上发表论文20余篇,主持省、市级自然科学基金项目各1项。

LI Cunhua was born in 1963. He received his Ph.D. degree in computer applications from Southeast University in 2004. Now he is a professor at Huaihai Institute of Technology. His research interests include data mining, artificial intelligence and image processing, etc.

李存华(1963—),男,江苏徐州人,2004年于东南大学计算机应用专业获得博士学位,现为淮海工学院教授,主要研究领域为数据挖掘,人工智能,图像处理等。在国内外期刊和会议上发表论文100余篇,主持国家、省、市级项目9项。

DAI Hongwei was born in 1975. He received his Ph.D. degree in system science from the University of Toyama in 2007. Now he is an associate professor at Huaihai Institute of Technology. His research interests include artificial intelligence and data mining, etc.

戴红伟(1975—),男,河南新郑人,2007年于日本富山大学获得博士学位,现为淮海工学院副教授,主要研究领域为人工智能,数据挖掘等。在国内外期刊和会议上发表论文20余篇。

LIU Zongtian was born in 1946. He received his M.S. degree in software engineering from Beijing University of Aeronautics and Astronautics in 1982. Now he is a professor and Ph.D. supervisor at Shanghai University. His research interests include software engineering and artificial intelligence, etc.

刘宗田(1946—),男,江苏临沂人,1982年于北京航空航天大学软件工程专业获得硕士学位,现为上海大学教授、博士生导师,主要研究领域为软件工程,人工智能等。在国内外期刊和会议上发表论文300余篇,主持国家基金项目4项。