面向舆情事件的子话题标签生成模型 ET-TAG

周 楠[1),2),3)] 杜 攀[1),2)] 靳小龙[1),2)] 刘 悦[1),2)] 程学旗[1),2)]

[1)](中国科学院网络数据科学与技术重点实验室 北京 100190)

[2)](中国科学院计算技术研究所 北京 100190)

[3)](中国科学院大学 北京 100049)

摘 要 关于舆情事件的新闻数据是纷繁复杂的.即便是关于同一舆情事件的新闻数据,往往包含有不同的子话题(事件的不同侧面).因此,如何生成能够准确描述事件子话题含义的标签对深入分析舆情事件(包括掌握事件热点、监测发展走向等)具有重要意义.事件子话题标签的生成通常包括两个关键步骤:首先发现子话题,然后依据每个子话题的关键词或文档内容生成描述该子话题的有效标签.传统方法在发现话题时多采用聚类或分类的方法,它们将同一个话题的文档整合到一个簇中.然而,由于隶属同一事件的文档具有很强的相似性,现有方法难以度量他们之间的距离,因此无法应用于发现事件子话题这一任务.此外,在为子话题生成标签时,传统的方法通常通过抽取来实现.此类方法所生成标签的准确性无法保证.为此,该文提出了一种基于 PLSA with Background Language 并结合关键词聚类发现事件内部子话题,进而基于维基百科等知识库生成事件子话题标签的模型 ET-TAG.在多类舆情事件数据集上的实验结果表明,ET-TAG 算法相比 *K*-means 和 LDA 等已有子话题发现方法具有更好的性能;从子话题标签生成角度而言,ET-TAG 生成的标签相对于传统方法也具有更好的准确性和概括性.该文最后将 ET-TAG 算法生成的子话题标签用于事件的对比和追踪,结果表明通过子话题标签可以发现事件共性,并反映事件子话题热度的变化趋势.

关键词 子话题发现; PLSA with Background Language;关键词聚类;子话题标签生成

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2018.01490

ET-TAG: A Tag Generation Model for the Sub-Topics of Public Opinion Events

ZHOU Nan[1),2),3)] DU Pan[1),2)] JIN Xiao-Long[1),2)] LIU Yue[1),2)] CHENG Xue-Qi[1),2)]

[1)](*CAS Key Laboratory of Network Data Science & Technology*, *Beijing* 100190)

[2)](*Institute of Computing Technology*, *Chinese Academy of Sciences*, *Beijing* 100190)

[3)](*University of Chinese Academy of Sciences*, *Beijing* 100049)

Abstract The public opinion system is a system to monitor the trend of public opinion on the Web. Through the public opinion system, we can understand hot spots on the Web and track their trends. Events are the focus of the public opinion system. News data about public opinion events are very complicated. Even for the data about the same event, it often contains different sub-topics (different perspective of the event). The sub-topics of an event can reflect its different aspects. For example, in the event of an earthquake, sub-topics include earthquake details, rescue work, post-disaster reconstruction, and so on. These sub-topics not only embody different aspects of the event, but also reflect the hot spots that public opinion may concern about. Tags of events sub-topics can be regarded as the attributes of events, which can help us to describe and comprehensively understand the events. Through sub-topics, we can compare the similarities and

收稿日期:2016-11-21;在线出版日期:2017-10-03.本课题得到国家自然科学基金(61572473,61472400)和国家青年科学基金(61303156)资助.**周 楠**,男,1991 年生,硕士,主要研究方向为数据挖掘、自然语言处理、机器学习. E-mail: zhounan@software.ict.ac.cn. **杜 攀**,男,1981 年生,博士,助理研究员,中国计算机学会(CCF)会员,主要研究方向为网络数据挖掘、机器学习、社交网络. **靳小龙**,男,1976 年生,博士,副研究员,博士生导师,中国计算机学会(CCF)会员,主要研究方向为社会计算、多智能体系统、性能建模与评估. **刘 悦**,女,1971 年生,博士,副研究员,中国计算机学会(CCF)会员,主要研究方向为信息检索、网络数据挖掘. **程学旗**,男,1971 年生,博士,研究员,博士生导师,中国计算机学会(CCF)会员,主要研究领域为网络科学、网络数据检索与挖掘.

differences between different events, and the sub-topic tags in a certain period of time can reflect changes in public opinion for the spots of events. It is significance to detect sub-topics of events and generate accurate sub-topic tags for public opinion system. It usually contains two major steps to generate the tags of sub-topics of a public opinion event: It first discovers sub-topics and then generates effective tags for them based on their corresponding keywords and documents. Existing methods for discovering topics or sub-topics are usually based on clustering or classification, which put the documents about the same topic into the same cluster. However, as the documents about the same event are similar to each other, it is very difficult for existing methods to measure the distance between these documents and thus they cannot effectively differentiate the sub-topics in the same event. There are a lot of high frequency background words in each document, how to ensure the diversity of sub-topics is a big problem. In addition, traditional methods often employ an extraction based manner to generate sub-topics' tags, where the accuracy of the tags cannot be guaranteed. And it is difficult to ensure the intelligibility of the generated tags. For overcoming such problems, this paper proposes an ET-TAG model, which uses PLSA-BLM to discover sub-topic keywords, KL divergence to merge similar sub-topics, and then utilizes co-occurrence relations to update sub-topic keywords. Based on the sub-topic keywords, the external knowledge base is used to generate the corresponding tags for each sub-topic. ET-TAG has higher accuracy when generating sub-topic tags, ET-TAG performs much better. Furthermore, the tags generated by ET-TAG are more accurate and summary. Finally, the tags generated by Experiments on Sogou news corpus and specific multi-category public opinion events corpus can prove that ET-TAG has obvious advantages compared with traditional methods(including *K*-means and LDA) in sub-topic discovery. It has higher accuracy when generating sub-topic tags. ET-TAG is used to compare and track events, which shows that sub-topic tags may help find the common points between different events and reflect the heat trends of the sub-topics of events.

Keywords　Sub-topics detection; PLSA with Background Language Model; Key words clustering; sub-topic tag generation

1　引　言

舆情系统是监测和分析网络舆情热点、趋势的系统，通过舆情系统可以了解网络舆论的关注热点，及其变化趋势. 在实际的舆情系统中，事件往往是关注的重点，舆情事件指的是民众广泛关注并容易引起舆论强烈反响的事件，例如突发事件或自然灾害. 特定的舆情事件往往伴随着不同的子话题，每一个子话题都在描述事件的不同侧面. 例如有关地震的事件包含的子话题有：地震详情、救援工作、灾后重建、社会反响以及责任追究等. 这些事件的子话题既是事件内容的不同角度描述，也反映了公众舆论可能的关注热点，对于舆情事件的分析有很大帮助. 着眼于事件的子话题使我们可以更好地跟踪事件的发展与变化，并对舆论的关注热点进行研究. 子话题的标签可以描述子话题的核心含义，便于我们直观地了解事件所涉及的方方面面. 通过这些标签我们可以清晰、简明地了解事件，可以对比不同事件的异同，也可以跟踪同一事件子话题的变化，因此子话题标签可以为事件的分析提供极大便利.

舆情事件子话题标签生成这一任务，有其特殊之处，通常包含两个关键步骤：首先要发现事件中的子话题，其次基于子话题的关键词或子话题内部的文档内容生成标签. 发现子话题这一步骤的难点在于很难保证子话题的差异性，而标签生成的难点体现在如何提升标签的准确性.

目前对于事件话题发现相关的研究多从文档角度或关键词的角度展开分析，文档角度的分析多采取聚类、分类等方法，基于文档相似度的计算. 从词的角度出发，可利用信号处理或主题模型等手段发现体现话题含义的关键词. 传统的方法在分析舆情事件子话题时遇到的困难是：描述同一事件的文档往往十分相似，这样会直接影响分类或聚类的效果.

现有方法在话题标签生成时多采用无监督的方式,基于话题关键词或者关键短语抽取,这样生成的标签很难准确地概括话题,在分析事件子话题时不利于判断子话题的质量和相关性.综上,传统的方法无法在分析舆情事件子话题并生成标签这一任务中克服以上两个难点.

本文提出的面向舆情事件的子话题标签生成模型(ET-TAG)将围绕子话题标签生成的两大难点展开.首先利用 PLSA with Background Language (PLSA-BLM)更好地发现不同子话题的高频备选关键词;在此基础上,利用词与词之间的共现关系更新关键词词组从而得到最终高质量的子话题关键词;最后,依据这些关键词并结合舆情事件的外部知识库自动生成子话题标签.

本文第 2 节介绍常见话题发现、话题标签生成的相关工作;第 3 节描述 ET-TAG 模型的主要模块;第 4 节对所提出的模型进行对比实验验证;第 5 节将 ET-TAG 模型生成的事件子话题标签应用于事件对比和子话题趋势研究;第 6 节对全文进行总结.

2 相关工作

总体而言,生成话题标签的基础是发现话题,话题发现的质量直接决定了标签生成的效果.关于话题发现的研究可大致从两个角度入手:文档角度和词的角度.

从文档角度出发,传统的话题发现与追踪技术(TDT)[1-2]可用来发现事件和话题,多采用聚类和分类的方法,通常采用的算法有:局部敏感哈希、增量聚类、分类算法等.

目前对于话题的研究很多基于微博或 Twitter 等社交媒体语料[3-4],也可以综合利用多种自媒体数据来源[5].针对流式自媒体数据可以采用局部敏感哈希[6]可以大大节省内存提升效率,但是不同的哈希函数基于不同的距离度量方法,采用最近邻的思路可能会漏掉大量相关文档,同时无法保证发现话题的可理解性.

基于文档聚类[7]的话题发现代表性方法 single pass clustering algorithm[8-9],将文档映射到特征空间,计算最新文档与之前文档的相似度,如果与之前任何一篇文档的相似度均未超过阈值,则视为新话题文档.然而对于同一事件的报道相似度很高,聚类算法无法将不同的事件"侧面"加以区分,而且对于利用 K-means 聚类的方法而言,参数 k 的确定十分困难,初始点的选择极大地影响话题的发现效果.同样可以利用层次聚类实现话题发现[10],利用 LSA 以及 LDA 提取文档特征,基于层次聚类发现新话题.然而该方法仍然基于文档相似性的比较,在处理同一事件的文档时效果不佳.

基于文档分类的话题发现算法[11-12]用分类器来判断文档是否属于特定话题,文档的特征包括:统计特征、关键词特征和上下文特征.分类的方法需要复杂的特征工程和大量的标注数据,而且话题必须事先定义好,因此这类方法只能适用于特定领域,对于模型的可扩展性带来一定的隐患.

从词的角度发现话题,大致有以下几种思路:借鉴信号处理的方法研究词频的变化、利用主题模型建模、对词序列建模、构建词图挖掘子话题等.

新话题的出现往往伴随着某些关键词的词频突变,因此可以借鉴信号处理的思路来研究词频的变化,利用小波变换可以为每个词建模[13],捕捉词频随时间的推移而可能发生的突变,采用卡尔曼滤波比较词频信号的相似性,事件话题发现则采用基于图划分的聚类方法.但是此类方法发现的关键词很难保证含义具有关联性,话题难以被人们直观理解.此外,利用主题模型 LDA 发现话题也是一种常用的手段[14],但是 LDA 需要对文档进行预处理,并且其把 LDA 的结果直接作为话题结果,话题质量不高,可理解性不强.

基于 HMM[15]的话题发现将原问题转化为词粒度的序列标注问题,为文档中的每一个词生成一个 topic 标签,根据预测出的 topic 序列生成最终的文档话题.利用 RNN[16]可以发现文本中蕴含的话题关键词和事件的触发词.然而,以上两类工作标注的工作量巨大并且最终的话题判定有一定的不确定性,需要大量的启发式规则,缺乏普适性.

利用微博特有的 hashtag 和发微博所在地区的地理信息也可以作为话题、事件发现的重要依据[17],但是这种方法局限于微博的特定格式,难以推广到新闻、论坛等语料.利用词的共现可以构建词图结构[18],在词图的基础上可以发现具有紧密共现关系的词,这些词可以作为话题的代表.但是此类方法很难保证话题的差异性,不能满足舆情事件的分析需求.利用图结构可以度量话题之间的关系,贝叶斯网络可以构建出话题之间的因果关系[19],话题的发现可以采用社区发现的相关算法.这类方法在图的构建上存在困难,且同样很难总结出话题的含义.

基于发现的话题生成相应的标签,现有方法采

用无监督的思想，大致可分为三类. 第一类方法是在 LDA 的基础上为每个词生成不同的权重[20]，利用权重对 LDA 发现的主题词进行调整，或对 LDA 发现的关键词进行组合[21]，在各种词的组合中抽取短语. 第二类方法利用词性信息对关键短语进行抽取[22]，基于这些短语和词表示话题. 第三类则引入了语义信息[23]，将文档中的词分为名词、形容词、动词三类，可以表达话题涉及到的事物、行为和描述性信息. 但是这类无监督的标签生成方法在处理舆情事件时生成的子话题标签准确性无法保证，一些短语表意不完整，这类标签无法实现对事件发展的跟踪和追溯.

综上，传统的话题发现方法和标签抽取方法没有考虑发现话题的合理性，无法对发现的话题之间的关系进行比较分析，出现重复话题的可能性较大，生成的话题标签质量不高. 因此本文在发现舆情事件子话题的基础上结合外部知识库为每一个子话题生成相应的标签. 基于子话题标签，可以实现子话题的演变跟踪、事件共性比较等后续工作，可以提高子话题的实际应用价值.

3　ET-TAG 模型

为了解决传统方法的事件子话题区分度差与子话题标签不准确的问题，本文提出了 ET-TAG 模型. 事件子话题区分度差的原因主要在于描述同一事件的文档内部有较强的相似性，大量的背景词频繁地出现在各个文档中，这些词大大降低了子话题之间的区分度. 为此，ET-TAG 模型采用 PLSA-BLM，通过在传统的 PLSA 的基础上引入背景语言模型，可以降低背景词对子话题的干扰. 此外，传统方法多采用无监督抽取的思路抽取子话题标签，这导致子话题标签的准确率较低. 如果将一些关键词和短语直接当作标签，则无法保证这些词或短语可以很好地概括子话题. 在 ET-TAG 模型中，采用有监督的思想，从外部的知识库引入特定类别舆情事件的概念体系，将其当作事件内部子话题的标签可以提高标签的可理解性.

ET-TAG 模型的目标是在给定一个舆情事件的相关新闻的前提下，发现描述该事件不同侧面的子话题，并为每个子话题生成描述其核心含义的子话题标签. 模型的核心环节如下：(1) 基于 PLSA-BLM 发现子话题以及相应的关键词；(2) 合并高相似度的子话题；(3) 基于共现关系利用并查集实现同一子话题内部的关键词聚类，每一个词簇可表示最终得到的子话题；(4) 根据词簇结合外部知识库，生成每个子话题的含义标签，模型流程如图 1 所示. 利用本模型生成的子话题含义标签可对子话题深入研究.

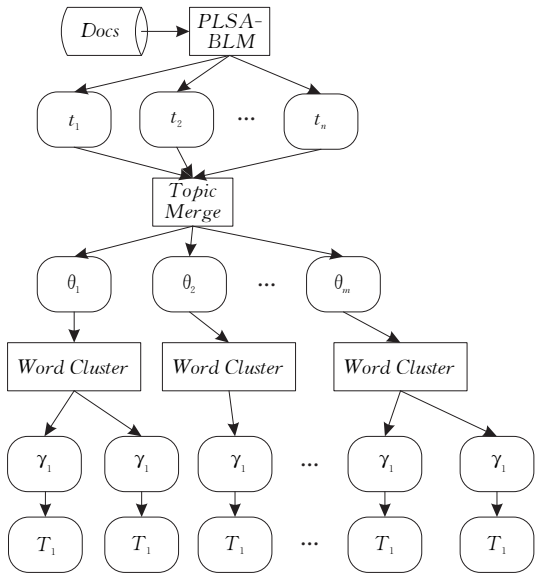


图 1　ET-TAG 模型示意图

3.1　基于 PLSA-BLM 的子话题发现

利用主题模型可以从词分布的角度发现事件的子话题，为了提升子话题之间的区分度需要去掉在多篇文档中均出现多次的背景词，传统的 LSI、LDA 等方法需要对文档进行预处理，并且无法有效去除背景词，因此本文采用 PLSA-BLM[24] 主题模型来解决此类问题.

主题模型是针对文档中的隐含主题建模的方法. PLSA 是基于统计的主题模型，参数估计采用 EM 算法，概率图模型如图 2 所示

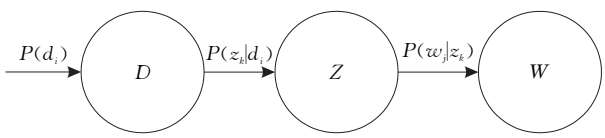


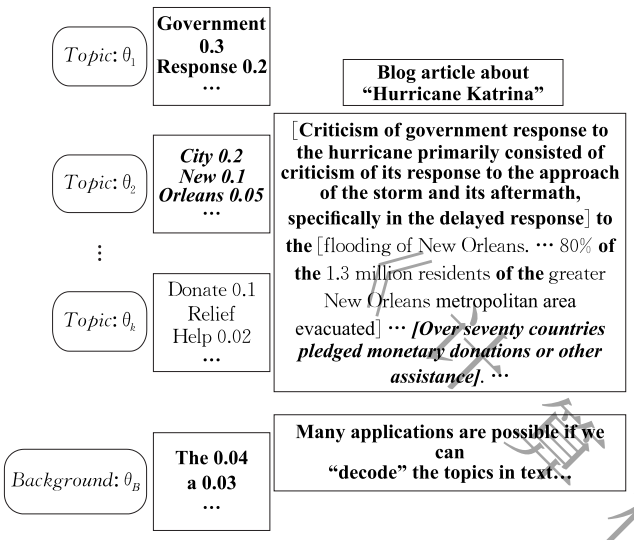
图 2　PLSA 概率图模型表示

$P(d_i)$是选中 d_i这篇文档的概率，$P(z_k|d_i)$是为 d_i这篇文档赋予主题 z_k的概率，而 $P(w_j|z_k)$是从 z_k这个主题中生成词 w_j的概率. 其中 $P(z_k|d_i)$. 和 $P(w_j|z_k)$均满足多项式分布. 通过 EM 算法可以求解以上参数.

传统的 PLSA 算法采用极大似然的方法进行参数估计，但是描述同一事件的文档存在大量背景词，这些背景词个文档中均有出现，这就导致发现的主题词区分度不大，无法满足利用主题词来发现事

件的子话题的需求.因此本文采用 PLSA-BLM,对这一问题加以改进.

PLSA-BLM 的核心思想是在传统 PLSA 的基础上引入背景语言模型.不妨假设原来的主题为 θ_1,θ_2,…,θ_k,在此基础上引入背景主题 θ_B,引入背景主题之后,可以将大量出现的高 df 词、停用词归纳到 θ_B 背景主题中,这部分词对于主题发现将不会有影响.模型最终效果如图 3 所示.



注:不同的字体代表不同的主题,文档中的每个词都被划分到一个特定的主题中.

图 3 PLSA-BLM 示意图

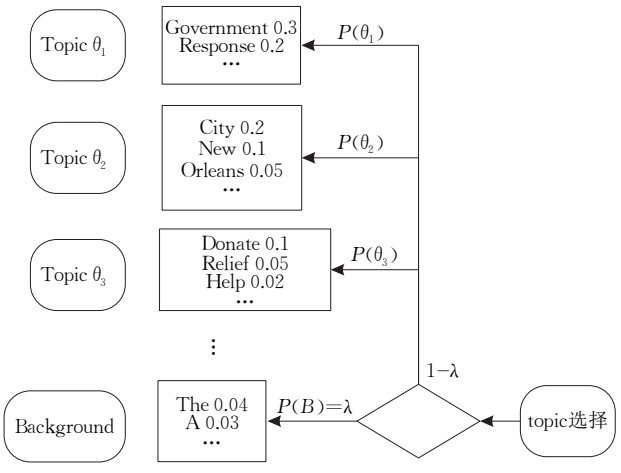


图 4 PLSA-BLM 文档生成过程示意图

在传统 PLSA 的基础上,对公式进行如下修改,$\pi_{d,j}$ 为文档 d 属于主题 j 的概率.$P(w|\theta_j)$ 为词 w 在 θ_j 中的概率.则生成文档 d 中词 w 的概率为

$$P_d(w)=\lambda_B p(w\mid\theta_B)+(1-\lambda_B)\sum_{j=1}\pi_{d,j}p(w\mid\theta_j) \tag{1}$$

生成文档 d 的概率为

$$\log P(d)=\sum_{w\in V}c(w,d)\log[\lambda_B p(w|\theta_B)+(1-\lambda_B)\sum_{j=1}\pi_{d,j}p(w|\theta_j)] \tag{2}$$

生成整个文档集合的概率为

$$\log P(C|\Lambda)=\sum_{d\in D}\sum_{w\in V}c(w,d)\log[\lambda_B p(w|\theta_B)+(1-\lambda_B)\sum_{j=1}\pi_{d,j}p(w|\theta_j)] \tag{3}$$

其中 $\Lambda=(\{\pi_{d,j}\},\{\theta_j\})$,$j=1,\cdots,k$.

采用 EM 算法对模型参数进行估计过程如下:

E-step:

$$P(z_{d,w}=j)=\frac{\pi_{d,j}^{(n)}P^{(n)}(w\mid\theta_j)}{\sum_{j'}^{k}\pi_{d,j'}^{(n)}P^{(n)}(w\mid\theta_{j'})} \tag{4}$$

$$P(z_{d,w}=B)=\frac{\lambda_B P(w\mid\theta_B)}{\lambda_B P(w\mid\theta_B)+(1-\lambda_B)\sum_{j}^{k}\pi_{d,j}^{(n)}P^{(n)}(w\mid\theta_j)} \tag{5}$$

M-step:

$$\pi_{d,j}^{(n+1)}=\frac{\sum_{w\in V}c(w,d)(1-p(z_{d,w,B}=B))p(z_{d,w}=j)}{\sum_{w'\in V}\sum_{d\in C}c(w'-d)(1-p(z_{d,w'}=B))p(z_{d,w'}=j)} \tag{6}$$

$$P^{(n+1)}(w\mid\theta_j)=\frac{\sum_{d\in C}c(w,d)(1-p(z_{d,w}=B))p(z_{d,w=j})}{\sum_{w'\in V}\sum_{d\in C}c(w',d)(1-p(z_{d,w'}=B))p(z_{d,w'}=j')} \tag{7}$$

文档生成的过程如图 4 所示,利用 PLSA-BLM 可以得到每个主题的词频分布.这些主题可视为初步发现的题.至此,第一步的工作已经完成.

3.2 基于 KL 散度的相似子话题合并

利用 PLSA-BLM,可以在每个子话题下得到不同的词频分布,但是得到的分布有可能有很强的近似性.因此发现相似的子话题并对其进行合并可以去除冗余,改善子话题发现的效果.

不同子话题的词频分布会有差异,而区分度越大的子话题,词频分布差异也就越大.因此可以利用比较两个子话题词频分布的差异来度量子话题之间的差别.

KL 散度(Kullback-Leibler Divergence),也称为相对熵,可以衡量相同事件空间下两个分布的差异.根据信息论中的定义,KL 散度的物理意义是在相同的事件空间下,概率 $P(x)$ 若用 $Q(x)$ 编码,平均每个符号编码长度增加了多少比特.用 $D(P\|Q)$ 表

示 KL 散度，计算公式如下：

$$D(P\|Q)=\sum_{x\in X}P(x)\log\frac{P(x)}{Q(x)} \tag{8}$$

利用 KL 散度可以比较不同子话题下词频分布的差异，考虑到 KL 散度不具有对称性，可采用以下公式计算分布之间的距离：

$$Diff(P(\theta_i),P(\theta_j))=$$
$$(D(P(\theta_i)\|P(\theta_j))+D(P(\theta_j)\|P(\theta_i)))/2 \tag{9}$$

KL 散度具有非负性，如果两个分布完全相同则 KL 散度为 0. $P(\theta_i)$和 $P(\theta_j)$代表子话题 θ_i和θ_j的词频分布. 对得到的子话题进行两两比较，计算分布差异度. 如果 $Diff$ 小于给定阈值，则子话题 θ_i和 θ_j 应当合并. 在子话题合并之后，将每个子话题下的关键词按照词频降序排列，词频越高说明这个词在当前子话题下越有代表性. 截取频率最高的 k 个词，作为合并后的子话题的关键词.

3.3　基于关键词共现更新子话题

经过子话题合并，可以得到新子话题内部的 top$k$ 关键词. 但是关键词之间关联度往往不高，甚至关键词内部可能蕴含多种不同的语义. 因此我们有必要对这些词进行进一步的处理. 高相关的词往往具有明显的共现关系，以句子为单位统计每个子话题关键词内部的共现关系，构建关键词共现矩阵示意图如图 5 所示.

	$W1$	$W2$	$W3$	$W4$
$W1$	—	2	0	3
$W2$	2	—	2	1
$W3$	3	1	—	2
$W4$	1	2	2	—

图 5　关键词共现矩阵

依据关键词共现矩阵，可以构建一个词图. 设定阈值 θ，$c(w1,w2)$和 $w2$ 在固定窗口内的共现次数果 $c(w1,w2)$大于 θ，$w1$ 和 $w2$ 可视为有连边，构造得到的示意图如图 5 所示，利用并将全部有连边的词聚到一个簇内. 簇内的词关系，而簇与簇之间共现关系不明显过此步骤，得到一系列的词簇. 如果词簇的大小过 2，则将其舍弃(少于三个关键词不足以表晰的语义). 图 6～图 7 展示了运行并查集之后得到的词簇示意图，图中的虚线框 A、B 和 C 分表不同的词簇，由于簇 C 内部的关键词个数较所被舍弃. 词簇内部的语义关联度更为紧密，因此这些高质量的词簇可作为子话题的最终表现形式簇反向检索文档，可以得到属于每个子话题下的文档.

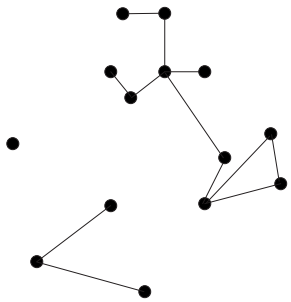


图 6　子话题内部关键词共现关系

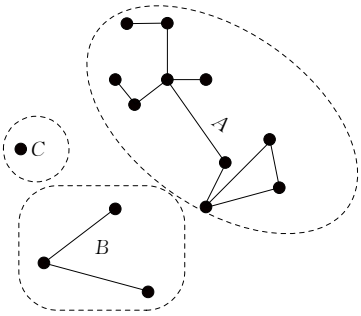


图 7　处理后的子话题关键词簇

3.4　子话题标签生成

目前为止，我们已经得到了若干舆情事件子话题及相应的关键词，仅仅将子话题的表示视为一个简单的词袋，不便于之后的分析，也难以研究事件子话题的演变. 为了解决这些困难，本文希望在子话题关键词的基础上能提炼出更为简洁、准确的标签. 通过子话题各自不同的标签，人们可以方便地获取子话题的核心含义，对舆情事件可能涉及到的各个侧面有直观的了解.

子话题标签应该具有高度的抽象性，表述子话题核心的含义，而同类舆情事件的子话题往往有一定的相似性，例如不同的地震事件新闻都会涉及地震详情的描述，灾后损失，救援情况以及后续报道等等. 考虑到同类事件的相似性，我们可以为它们设计一组子话题标签，标签的设计参考维基百科等事件知识库. 在获得各个子话题关键词的基础上，接下来的工作是根据这些子话题关键词预测一个具体的子话题归属与哪一个标签.

判断子话题的标签需要一些包含舆情事件的知识库，这些知识库涵盖了所在领域的各个方面的信息，例如地震类别的事件知识库会涉及历史上国内外著名的地震事件，以及关于震中、震级、损失等方面的报道，这些不同的方面都是舆情事件内部的概念体系. 依据知识库可对同类事件不同侧面的信息进行汇总，生成子话题标签的过程实质上是建立子话题关键词到舆情知识库事件概念体系的映射关

系,来源于外部知识库的概念体系就是事件子话题的标签集合.本文采用的是维基百科知识库及网易新闻知识库,在多个事件类别上整理相关事件的文档,归纳概念体系作为预定义的子话题标签.

对知识库数据按内部概念进行整理、分词预处理以计算出知识库中的每一个词属于特定子话题标签(概念体系)的概率 $P(t_k|w)$.然而有一部分词在知识库中极少出现,这些低频词不利于子话题标签的推断,考虑到以上因素,可以计算知识库中的词对每一个子话题标签的贡献值,定义公式如下:

$$score(t_k,w_i)=P(t_k|w_i)\log(tf(w_i)) \quad (10)$$

子话题标签集合为 $T=\{t_1,t_2,\cdots,t_k\}$,$S(t,\theta_k)$为子话题 θ_k在标签 t 上的得分,定义如下

$$S(t,\theta_k)=\sum_{w\in\theta_k} score(t,w) \quad (11)$$

根据式(10)、(11)可以计算出子话题所属的标签,算法流程如下算法 1.

算法 1. 话题标签生成算法.

输入:$P(t|w)$,tf, T, θ

输出:$records$

```
1. records={}
2 FOR θ_t∈θ DO
3.    maxScore=0
4.    t_max=NULL
5.    FOR t_j∈T DO
6.      S(t_j,θ_i)=0
7.      FOR w_k∈θ_i DO
8.         score(t_j, θ_i)=P(t_j|w_k)log(tf(w_k))
9.         S(t_j,θ_i)+=score(t_j,θ_i)
10.      END FOR
11.      IF S(t_j,θ_i)>maxScore THEN
12.         maxScore=S(t_j,θ_i)
13.         t_max=t_j
14.      END IF
15.   END FOR
16.   records[θ_t]=t_max
17. END FOR
18. RETURN record
```

算法的输入 $P(t|w)$代表词 w 属于标签 t 的概率,tf 代表词表中各个词的全局词频,T 代表子话题的标签集合,θ 代表子话题集合.输出 $record$ 表示子话题和子话题标签的映射表.代码中的三重循环分别遍历子话题、子话题标签、子话题中的关键词,计算关键词在特定标签下的得分(8 行),累加到标签总得分中(9 行),选取得分最高的标签,赋值给当前的子话题(12~13 行).全部循环结束后可以得到所有子话题对应的子话题标签.

常见舆情事件的新闻语料行文比较规范,同类别新闻的描述类关键词十分相近,因此知识库中的词汇足以满足子话题标签预测时的需求.在给定同一舆情事件的新闻语料后,通过本文提供的方法可以结合维基百科等外部知识库自动生成舆情事件的子话题标签.

4 实验评估

在本节中,我们利用真实的新闻事件语料对算法的效果进行验证.验证可分为两个部分,首先对比 ET-TAG 与传统方法在事件子话题发现上的效果(搜狗新闻语料).其次验证子话题合并对效果的影响.最后利用人工对子话题标签的标注验证事件子话题标签预测的准确性,目前的实验针对自然灾害和突发事件两大类,其中自然灾害包括地震、台风.突发事件包括爆炸、空难、集会.

4.1 实验数据集

实验数据主要有两个来源,自然灾害类和突发类的舆情事件语料,总计 6000 余篇文档.其中地震类的语料包含的事件有:尼泊尔地震、台湾花莲地震、新疆地震、鲁甸地震的后续报道等.台风类事件包含灿鸿台风和苏迪罗台风的相关新闻.爆炸类语料包含的事件包括:天津爆炸、淄博化工厂爆炸、四川达州瓦斯爆炸等.此外还有马航 MH370 飞机失事以及香港"占中"等事件.测试子话题发现的效果采用的是已经分好类别的搜狗新闻语料,总计 17 000 余篇文档 9000 余词汇,涉及文体、历史、社会新闻、教育健康等 10 个类别.

在生成子话题标签时需要的外部知识库来自维基百科相关专题以及网易新闻.基于外部知识库概念体系的子话题标签如下,地震类:详情、影响、救援、重建、纪念.台风类:位置、灾情、防护救援、影响.爆炸事故类:背景、事故现场、灾难救援、后果影响、社会反响、后续情况.空难类:灾情、搜救、赔偿及影响.政治集会类:背景、影响、主体、后续情况.

4.2 评估标准

本文工作分为两部分,首先,根据原始新闻语料无监督地生成子话题和关键词.其次根据关键词和外部知识库生成子话题的标签.因此,评估的两个层面分别是第一步子话题发现的效果和第二步标签生成的准确性.

ET-TAG 利用主题模型生成子话题,因此我们

借鉴聚类的评价思路，利用已标注新闻类别搜狗新闻语料，运行主题模型后每篇文档都会被归入概率最大的主题下，因此可计算兰德指数（*RI*）和标准化互信息（*NMI*）来评价效果.

兰德指数将聚类理解为一系列的决策过程，*TP* 表示将相似文档归入同一个簇中，*TN* 表示将不相似的文档归入不同簇，*FP* 表示将不相似的文档归入相同簇，*FN* 表示将两篇相似的文档归入不同的簇. *RI* 实际计算的是正确决策的比率，计算公式如下：

$$RI = \frac{TP + TN}{TP + TN + FN + FP} \tag{12}$$

标准化互信息的依据信息论的相关知识，可比较两种分布的吻合程度，它的定义如下：

$$NMI(\Omega, C) = \frac{I(\Omega, C)}{[H(\Omega) + H(C)]/2} \tag{13}$$

其中 I 是互信息，定义如下：

$$I(\Omega, C) = \sum_k \sum_j P(\omega_k \cap c_j) \log\left(\frac{P(\omega_k \cap c_j)}{P(\omega_k)P(c_j)}\right) \tag{14}$$

H 代表信息的熵，定义如下：

$$H(\Omega) = -\sum_k P(\omega_k) \log P(\omega_k) \tag{15}$$

$P(\omega_k)$ 表示文档属于簇 ω_k 的概率，$P(c_j)$ 表示文档属于标注类别 c_j 的概率，$P(\omega_k \cap c_j)$ 表示文档属于 $\omega_k \cap c_j$ 的概率.

在评价生成的子话题标签时，可借助人工标注一部分事件的子话题标签，通过计算预测结果相对于人工标注的准确率，评价其效果.

4.3　实验结果

K-means 是传统的聚类算法，在话题发现的相关研究中被广泛使用，LDA 是经典的主题模型，LDA 的结果通常作为话题分析的基础，因此二者具有较好的代表性，我们选择 *K*-means 和 LDA 作为 baseline，在生成子话题这一步骤中 ET-TAG 与传统基于 *K*-means 和 LDA 的 *RI* 和 *NMI* 作对比. 对比结果如表1所示，可见从文档按照话题聚类的效果来看，ET-TAG 更加适应于事件子话题发现这一任务.

表 1　不同算法在搜狗语料上子话题发现的比较

算法	*RI*	*NMI*
K-means	0.428	0.151
GibbsLDA	0.869	0.481
PLSA-BLM(λ_B=0.2)	0.611	0.323
PLSA-BLM(λ_B=0.5)	0.792	0.421
PLSA-BLM(λ_B=0.9)	0.875	0.467
PLSA-BLM+SubTopicMerge	**0.881**	**0.479**
ET-TAG	**0.934**	**0.847**

其中 PLSA-BLM 模型中的 λ_B 控制背景词出现的概率，λ_B 越大背景词出现的概率越低. 从实验结果可以看出 λ_B 的取值极大影响实验结果，去除背景词可显著提升子话题发现效果. 基于 PLSA-BLM 发现的多个子话题内部会有一定的相似性，从实验结果可知，利用 KL 散度合并相似的子话题可以进一步提升效果.

合并子话题之后，每个子话题内按照词频取前 top$k$ 个关键词，k 的取值将直接影响子话题的发现效果，在搜狗新闻语料上的实验结果如图 8 所示. 由实验结果可知，如果 k 取值较小，经过下一步词聚类得到的词簇个数极少，使子话题发现效果不佳. 随着 k 值得增大，效果逐步提升，但当 k 值继续增大时，关键词过多会使词簇个数增多，但差异性减少，同样会降低子话题的发现效果.

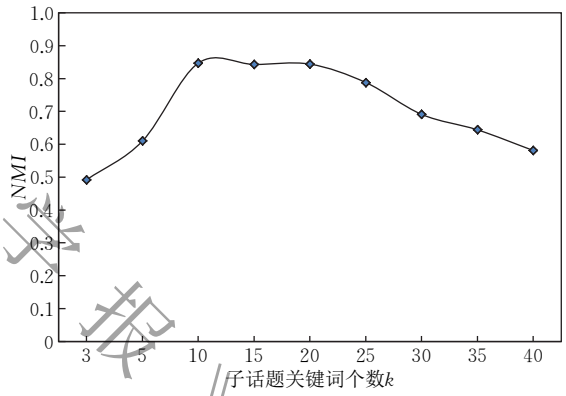


图 8　*NMI* 指数随子话题内部关键词个数的变化关系

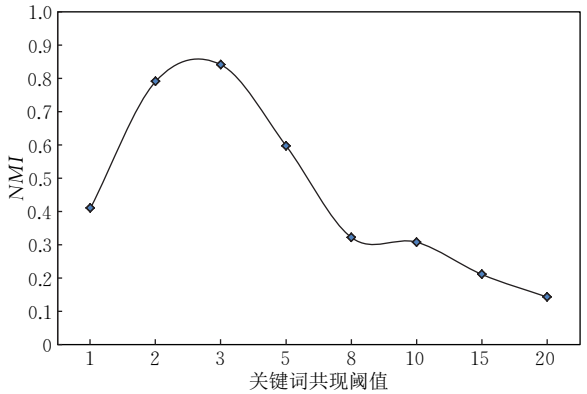


图 9　*NMI* 指数随 θ 取值的变化关系

在基于每个子话题下的关键词生成词簇的过程中需要指定关键词在指定窗口的共现阈值 θ，在实验中我们以句子长度为窗口，考察 θ 对 *NMI* 的影响，实验结果如图 9 所示. 从实验结果中不难发现，

当 θ 取值太小时，词簇个数少，词簇内部关键词个数多，这样导致每个词簇的核心含义并不明确. 而当 θ 太大时，对词之间的共现关系要求过高，这样每个词簇内关键词的个数将十分稀少，词簇的质量显著降低. 词簇内词的个数也会对最终子话题标签生成的准确性造成较大影响.

接下来以香港“占中”事件为例展示不同参数取值对子话题发现的影响，表 2 是 λ_B 取值为 0.5 时子话题发现的结果，表 3 是 λ_B 取值为 0.9 时子话题发现的结果，表 4 是加入子话题合并后的结果.

表 2 香港“占中”子话题发现结果(λ_B 取值为 0.5)

ID	子话题关键词
1	香港 政 改 立法会 基本法 普选 中 民 行政
2	香港「 港 中 反对派 政 改 议员 社会 普选
3	香港从军青年军队想死爱国解放军
4	香港 中央 中国 楼 一国两制 港人治港 港 落实 搞 自治
5	**美国美元中国经济资本走全球危机指数**
6	**美国美元霸权金融钱国家国际货币经济银行**
7	香港 台湾 港 国家 楼 中国 说 中 反 大陆

表 3 香港“占中”子话题发现结果(λ_B 取值为 0.9)

ID	子话题关键词
1	改 政 方案 学联 罗冠聪 黄之锋 秘书长 思潮 学生 6 月
2	港人治港 落实 楼 中央 普选 一国两制 选 37
3	泛 立法会 议员 主任 佔於 民 利益 态度
4	**分子 驻军 驻 部队 闯 挑衅 军营**
5	**分子 分裂 驻 日本 闯 部队 挑衅 岁 资料**
6	苑 学 声明 台 大学 历史
7	选择 加油 选举 水货 石油 游客 清

表 4 香港“占中”子话题发现结果(子话题合并)

ID	子话题关键词
1	改 政 方案 学联 罗冠聪 黄之锋 秘书长 思潮 学生 6 月
2	港人治港 落实 楼 中央 普选 一国两制 选 37
3	泛 立法会 议员 主任 佔於 民 利益 态度
4	分子 分裂 驻 日本 闯 部队 挑衅 岁 资料
5	选择 加油 选举 水货 石油 游客 清

观察以上实验结果，在 λ_B 取值为 0.5 时，大量的背景词例如：香港、美国等会出现在子话题关键词中，这些词会大大降低子话题发现的效果. 提高 λ_B 的取值，效果会有显著提升，但子话题之间会有相似结果，子话题合并后关键词质量可以进一步提升.

最后一步，借助人工标注的数据评价子话题标签预测的效果，采用准确率来度量，评估结果如表 5 所示.

由实验结果可知，对于典型的自然灾害和突发事件，子话题标签的预测结果较好，而对于香港“占中”和北京申冬奥会成功这样的事件效果不佳. 原因在于对于香港“占中”或申办冬奥会这类事件，事件本身比较复杂，可能涉及的子话题众多，同时外部知识库缺乏全面完整的概念体系整理，这直接影响了子话题标签的预测结果.

表 5 子话题标签预测准确率

类别	事件	标签预测准确率/%
地震类	尼泊尔地震	100.0
	鲁甸地震	83.3
	新疆地震	75.0
台风类	苏迪罗台风	75.0
	灿鸿台风	80.0
爆炸	天津爆炸	89.0
	淄博化工厂爆炸	80.0
	四川达州瓦斯爆炸	75.0
空难	MH370 飞机失事	100.0
集会	香港占中	65.0
运动会	北京申冬奥成功	66.0

表 6 是北京申请冬季奥运会成功的子话题标签的预测结果，这一事件本身具有复杂性，导致外部的知识库难以总结出有针对性的概念体系. 例如申冬奥成功后，对于企业界、城市规划甚至房价等都会造成影响，仅仅依据知识库本身的内容很难全面总结这些类别，对于此类复杂事件本文生成子话题标签方法尚有一定的局限性.

表 6 北京申冬奥会成功子话题标签预测结果

子话题关键词	标注标签	预测标签
造雪，人口，人工，能力，主办	准备工作	**其它**
健身，全民，体育，规划，强国，活动，中心，意义	意义	意义
陈述，邮票，设计，陈述，七匹狼，智慧申，纪念	影响	**其它**
元，旅游，水立方，楼盘，保护区，平方米，万元	影响	影响
企业，河北省，治理，赛区，高铁，交通，照明，张家口	准备工作	准备工作
新疆，钢材，评估，高速，机场，需求，地铁	影响	影响

表 7 和表 8 分别是尼泊尔地震和鲁甸地震的子话题标签预测结果，表 9 则是天津爆炸事件的子话题标签预测结果. 自然灾害和爆炸类事件子话题标题生成的准确率较高.

表 7 尼泊尔地震子话题标签预测结果

子话题关键词	标注标签	预测标签
26 日，里氏，救援，15 时，国际	救援	救援
吉隆，直升机，药品，输送，聂拉木县，吉隆县，西藏，震感，倒塌，日喀则	影响	影响
深度，震源，北纬，东经，20 千，14 时，11 分，28.2	详情	详情
30，寺庙，失踪，受损，程度，27 日，聂拉木	影响	影响
公路，交通，武警，路段，樟木口岸，部队，8 月	救援	救援
群众，18 时，各族，49	影响	影响

表 8 鲁甸地震子话题标签预测结果

子话题关键词	标注标签	预测标签
学校，活动，搭建，开学，负责	影响	救援
工作组，国务院，民政部，赶赴，国家	重建	重建
震源，断裂，历史，记录	详情	详情
遇难，8 月，昭通市	影响	影响
作出，人员，主席，妥善	救援	救援
队员，应急，平台，指挥部	救援	救援

表 9　天津爆炸子话题标签预测结果

子话题关键词	标注标签	预测标签
检察，犯罪，渎职，机关，高检，职务，依法	其它	背景
应急，生化，核心，国家，救援，15 日，第四	后续情况	后续情况
遇难，身份，确认，16 日，112，发现，9 时	灾难救援	灾难救援
消防，塘沽，开发区，11 时，一带，火光，冲天	后果影响	后果影响
44，住院，重症，包括，12，12 时，520，中午，66，13 日	事故现场	事故现场
所属，区域，17，国际，物流，中心，入院，32	背景	背景
特别，瑞海，人数，15 日，指挥部，104，获悉，8·12，火灾，公布	后果影响	后果影响
生产，会议，落实，部署，书记	后续情况	后续情况
周边，方圆，喷发，引发	事故现场	背景

综上，通过实验阶段的评测，本文提出的 ET-TAG 模型可以有效地发现事件的子话题，并为其生成准确的标签，生成的事件子话题标签可以为事件的进一步研究提供帮助.

5　子话题标签的应用

通常的事件、话题发现技术的目标通常停留在“发现话题”这一步骤，而子话题标签可以作为事件的一个重要属性帮助我们刻画事件、全面理解事件包含哪些侧面或分析角度，对事件的发展变化进行进一步分析，本节将对基于子话题标签的事件分析进行简单介绍.

事件最基本的属性包括时间、地点、涉及的人物等. 但是对于舆情事件这一特殊的事件类型，更值得关注的是公众舆论对于事件的反响和关注热点，仅仅提供最基本的事件要素难以满足舆情系统使用者的实际需求. 舆情事件的子话题标签代表了事件可能存在的不同的侧面，不同的侧面也对应了舆论不同的关注点，因此可以视为事件的“属性”.

子话题标签这种特殊的属性可以帮助我们全面深入地刻画事件、理解事件. 分析舆情事件又可以从两个角度入手. 一方面可以比较同类事件的子话题标签，研究公众对此类事件的关注热点是否有共性，或不同事件之间的差异点. 另一方面针对某个特定事件，可以从时间维度展开分析，研究事件子话题热度随着时间推移所产生的变化，热度的变化也反映出了舆论关注点的迁移.

以上两个角度的分析结果，可以为相关部门在面对突发事件所造成的舆论影响时所做出的处置决策提供依据.

在同类事件中，通过比较不同事件的子话题标签可以了解事件之间的异同点. 表 10～表 12 展示了四川、新疆两地的地震事件的基于子话题标签的对比.

表 10　四川、台湾地震事件子话题对比

事件	子话题关键词	共性标签
四川	冷木沟，城，村民，悬在，物源，下游	详情
台湾	位置，花莲，东北方	

表 11　四川、新疆地震子话题对比

事件	子话题关键词	共性标签
四川	广播，预警，启用，26，应急，两省，播出，实验室，安装	救援
新疆	公安局，民警，赶赴，公安机关，抢险救灾	

表 12　台湾、新疆地震子话题对比

事件	子话题关键词	共性标签
台湾	昨日，约，15.6，万	影响
新疆	馀，震，地震局，460，震区，专家，次，衰减，活跃，分析	

观察表 10～表 12 的事件对比结果，可以发现不同的地震事件的新闻报道都会涉及到的一些子话题：详情、救援、影响. 这些子话题可以体现出同类事件的共性，而事件子话题的共性也体现了事件之间的相似程度.

依据子话题标签不仅可以比较不同事件，也可以考察一段时间内舆论对于同一事件关注角度的变化以及事件本身的发展历程. 图 10 为天津爆炸事件各个子话题热度演变图，横轴为日期时间，纵轴为热度（子话题内部的文章总数），可以看到不同的子话题在报道时间上存在明显差异，不同子话题的热度也随着时间推移呈现出不同的姿态.

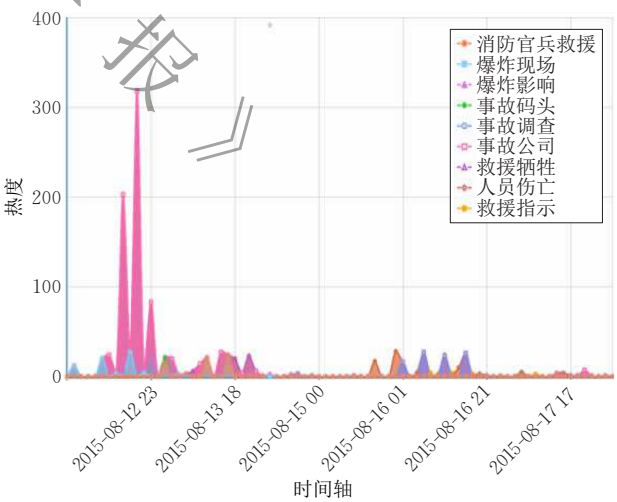


图 10　天津爆炸事件子话题热度变化图

图 11 反映的是标签为“爆炸现场”的子话题热度演化趋势，可以明显看出在事件的开始阶段舆论的关注点是爆炸的现场和危害程度.

随着时间的推移舆论关注热点从“爆炸现场”转变为“救援情况”，再接下来大家开始关注爆炸事故的人员伤亡情况. 最后事件尘埃落定，事故的调查结

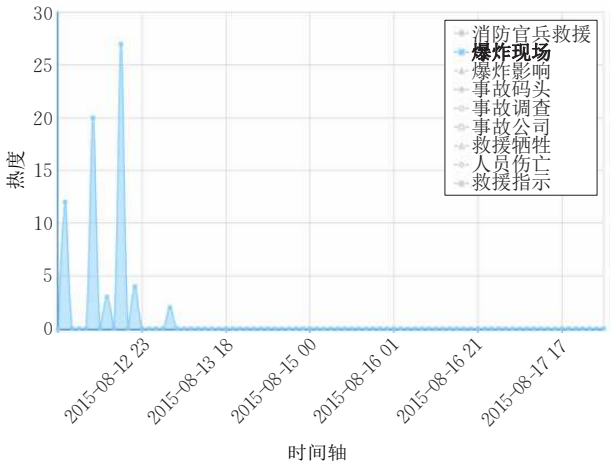


图 11 爆炸现场子话题演化趋势

果又称为关注热点.图 12～图 14 展示了以上过程中子话题热度的变化趋势.

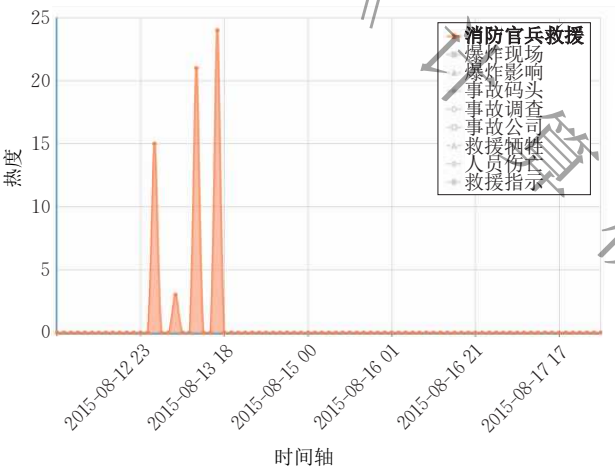


图 12 救援情况子话题演化趋势

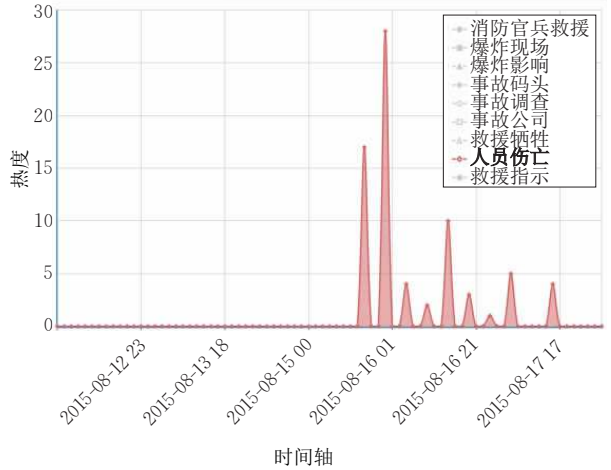


图 13 人员伤亡子话题演化趋势

同理,地震类别事件的子话题热度也随着时间变化呈现不同姿态,图 15 是新疆地震事件所有子话题的热度变化图,可以明显地看出不同子话题的热

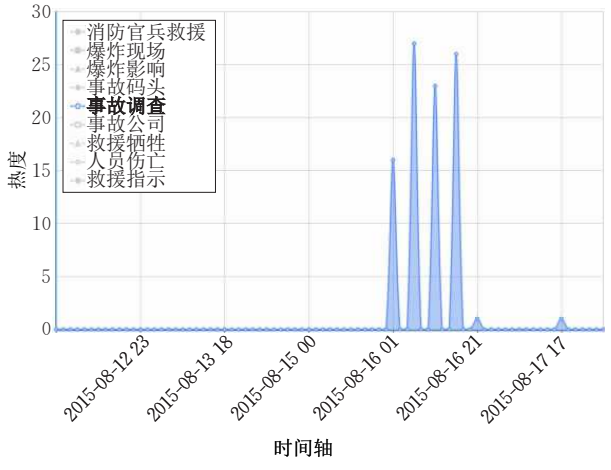


图 14 事故调查子话题演化趋势

度峰值往往出现在不同时间段,图 16～图 18 展示的是事件不同子话题的演化趋势,从图中可以明显地看出舆论的热点最初集中在震源位置的确定,之后开始密切关注地震造成的损失,当整个事件尘埃落定后,人们讨论更多的是灾后的安置措施.

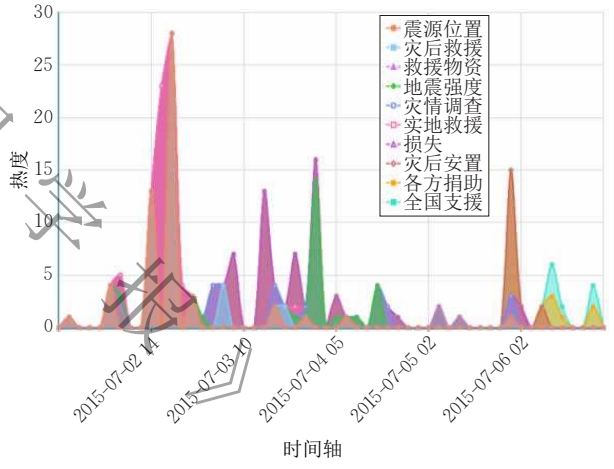


图 15 新疆地震事件子话题热度变化图

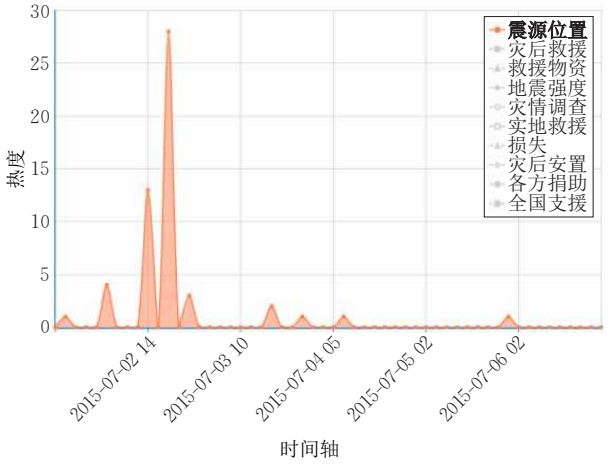


图 16 震源位置子话题演化趋势

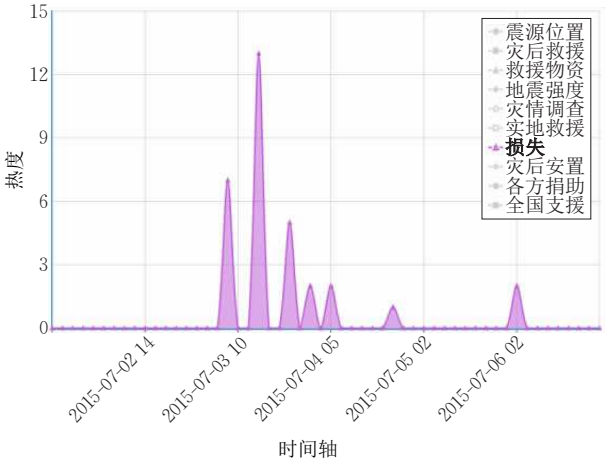


图 17　地震损失子话题演化趋势

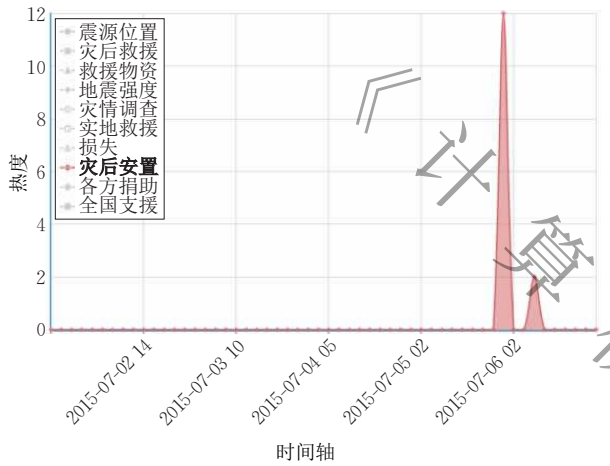


图 18　灾后安置子话题演化趋势

可见，通过将子话题标签和热度相结合，我们不仅可以清晰地了解舆情事件所涉及到的不同角度，还可以追踪舆论关注点的变化，这些都可以为针对舆情事件的进一步分析提供帮助.

综上所述，利用事件的子话题标签我们不仅可以简洁明了地了解同一事件的不同关注角度，还可以对比不同的事件、研究同一事件的变化趋势. 这些分析方法及分析结果对于舆情分析提供了帮助，体现了本文工作的意义所在.

6　总　结

舆情系统的重点是监测网路舆情的热点、趋势，并给出相关的分析. 其中针对突发事件以及自然灾害类的舆情事件分析是舆情系统的重要功能.

子话题指的是舆情事件内部的不同侧面或关注角度，发现事件的子话题可以帮助我们全面了解事件. 子话题标签是对子话题核心含义的概括，标签可以极大地提升子话题的可理解性. 通过舆情事件的子话题标签可以直观而全面地了解事件的不同侧面以及公众舆论的热点迁移.

舆情事件子话题的发现和标签生成是一类特殊的话题发现任务. 其特殊性体现在处理的文档都是关于同一事件的报道或描述，这导致文档之间的相似性较高，有大量的高频背景词出现在各个文档之中，背景词会影响子话题关键词的质量，使得关键词之间的差异减小. 如何保证发现的子话题有足够的差异性是一大难题.

子话题标签作为子话题核心含义的代表，可以帮助我们直观理解事件的不同侧面，如何能提升标签的可理解性也是一大难点.

当前对于话题发现的研究大致的思路包括：分类、聚类、话题关键词挖掘等. 然而以上工作通常需要度量文档之间的距离，描述相同事件的文档之间往往十分相似，这导致此类方法很难区分文档. 此外，传统方法对背景词缺乏针对性的处理，难以保证子话题之间有足够的差异性，无法应用在舆情事件子话题发现的任务中. 自动生成话题的标签，传统的做法的核心思路是基于无监督的关键词抽取，抽取出的关键词质量参差不齐，显然这种做法无法生成可理解性强的标签.

基于舆情事件分析的需求和研究现状，本文提出了 ET-TAG 模型，利用 PLSA-BLM 发现子话题关键词，利用 KL 散度合并相似的子话题，再结合词的共现关系更新子话题关键词. 在得到子话题关键词的基础上，结合外部知识库为每一个子话题生成对应的标签. 在搜狗新闻语料和具体的多类舆情事件语料上的实验可以证明在子话题发现这个角度，ET-TAG 相比传统方法有明显优势，在生成子话题标签时 ET-TAG 有较高的准确率.

利用生成的子话题标签可以进一步研究事件. 比较不同事件的子话题可以分析事件之间的异同总结同类事件的舆论关注重点，也可以分析一个具体事件在某一段时间内的子话题热度变化趋势. 舆情事件的分析结果可以为舆情系统的使用者提供更多的决策参考.

未来我们将在子话题发现阶段，在 PLSA-BLM 基础上融合其它方法，并且引入时间因素改善子话题关键词的质量. 生成子话题标签时，将外部概念体系和文档内容相结合，综合有监督方法和无监督方

法,在保证准确性、可理解性的基础上,生成更具体、更有特点的标签.

参 考 文 献

[1] Allan J. Topic detection and tracking: Event-based information organization. Springer Science & Business Media, Berlin, German: Springer, 2012

[2] He T, Qu G, Li S, et al. Semi-automatic hot event detection// Proceedings of the International Conference on Advanced Data Mining and Applications. Berlin, Germany, 2006: 1008-1016

[3] Aiello L M, Petkos G, Martin C, et al. Sensing trending topics in Twitter. IEEE Transactions on Multimedia, 2013, 15(6): 1268-1282

[4] Becker H, Naaman M, Gravano L. Beyond trending topics: Real-world event identification on twitter//Proceedings of the International AAAI Conference on Web and Social Media. Barcelona, Spain, 2011: 438-441

[5] Nguyen D T, Jung J E. Real-time event detection for online behavioral analysis of big social data. Future Generation Computer Systems, 2017, 66: 137-145

[6] Petkos G, Papadopoulos S, Kompatsiaris Y. Two-level message clustering for topic detection in twitter//Proceedings of the WWW. Seoul, Korea, 2014: 49-56

[7] Yang Y, Pierce T, Carbonell J. A study of retrospective and on-line event detection//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, 1998: 28-36

[8] Huang B, Yang Y, Mahmood A, et al. Microblog topic detection based on LDA model and single-pass clustering// Proceedings of the Rough Sets and Current Trends in Computing. Berlin, Germany, 2012: 166-171

[9] Allan J, Papka R, Lavrenko V. On-line new event detection and tracking//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, 1998: 37-45

[10] Yan X, Zhao H. Chinese microblog topic detection based on the latent semantic analysis and structural property. Journal of Networks, 2013, 8(4): 917-923

[11] Brants T, Chen F, Farahat A. A system for new event detection//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada, 2003: 330-337

[12] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: Real-time event detection by social sensors//Proceedings of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 851-860

[13] Weng J, Lee B S. Event detection in twitter//Proceedings of the International AAAI Conference on Web and Social Media. Barcelona, Spain, 2011, 11: 401-408

[14] Nallapati R, Feng A, Peng F, et al. Event threading within news topics//Proceedings of the 13th ACM International Conference on Information and Knowledge Management. Washington, USA, 2004: 446-453

[15] Hongeng S, Nevatia R. Large-scale event detection using semi-hidden Markov models//Proceedings of the 9th IEEE International Conference Louis. Missouri, USA, 2003: 1455-1462

[16] Ghaeini R, Fern X Z, Huang L, et al. Event nugget detection with forward-backward recurrent neural networks//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 369-384

[17] Cui A, Zhang M, Liu Y, et al. Discover breaking events with popular hashtags in twitter//Proceedings of the ACM International Conference on Information and Knowledge Management. New York, USA, 2012: 1794-1798

[18] Katragadda S, Virani S, Benton R, et al. Detection of event onset using Twitter//Proceedings of the 2016 International Joint Conference on Neural Networks. Vancouver, Canada, 2016: 1539-1546

[19] Drury B, Rocha C, Moura M F, et al. The extraction from news stories a causal topic centred Bayesian graph for sugarcane// Proceedings of the 20th International Database Engineering & Applications Symposium. Montreal, Canada, 2016: 364-369

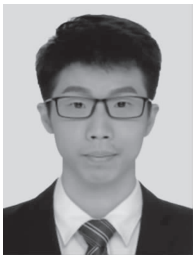
[20] Xu R, Ye L, Xu J. Reader's emotion prediction based on weighted latent Dirichlet allocation and multi-label k-nearest neighbor model. Journal of Computational Information Systems, 2013, 9(6): 2209-2216

[21] Johri N, Roth D, Tu Y. Experts' retrieval with multiword-enhanced author topic model//Proceedings of the NAACL HLT 2010 Workshop on Semantic Search. Association for Computational Linguistics. Uppsala, Sweden, 2010: 10-18

[22] Darling W M, Song F. Probabilistic topic and syntax modeling with part-of-speech LDA. arXiv preprint arXiv:1303.2826, 201

[23] Blei D M. Probabilistic topic models. Communications of the ACM, 2012, 55(4): 77-84

[24] Lu Y, Mei Q, Zhai C X. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. Information Retrieval, 2011, 14(2): 178-203



ZHOU Nan, born in 1991, M. S. His current research interests include data mining, natural language processing and machine learning.

DU Pan, born in 1981, Ph. D., assistant professor, His main research interests include Web search and mining, machine learning and social network.

JIN Xiao-Long, born in 1976, Ph. D., associate professor, Ph. D. supervisor. His research interests include social computing, multi-agent systems, performance modelling and evaluation.

LIU Yue, born in 1971, Ph. D., associate professor. Her main research interests include information retrieval and Web mining.

CHENG Xue-Qi, born in 1971, Ph. D., professor, Ph. D. supervisor. His research interests include network science, Web search & data mining.

Background

A public opinion system is a system to monitor and analyze the hot spots and the trend of public opinion on the Web. Through the public opinion system, we can understand hot spots on the Web and track their trends. Events are the focus of the public opinion system.

The sub-topics of an event can reflect its different aspects. For example, in the event of an earthquake, sub-topics include earthquake details, rescue work, post-disaster reconstruction, and so on. These sub-topics not only embody different aspects of the event, but also reflect the hot spots that public opinion may concern about. Sub-topics detection and labels generation of public opinion events is a special topic detection task. Its particularity is that the documents processed are the reports or description of the same event, which leads to the high similarity between the documents. There are a lot of high frequency background words in each document, how to ensure the diversity of sub-topics is a big problem.

Tags of events sub-topics can be regarded as the attributes of events, which can help us to describe and comprehensively understand the events. Through sub-topics, we can compare the similarities and differences between different events, The sub-topic tags in a certain period of time can reflect changes in public opinion for the spots of events. It is significance to detect sub-topics of events and generate accurate sub-topic tags for public opinion system.

Most of the traditional event discovery methods are based on the ideas of classification or clustering, which are highly dependent on feature engineering. Moreover, the events addressed by the traditional event discovery methods are often confined to a specific domain. This brings a great limitation to these existing methods, which makes it difficult to generalize them to events in other domains. In addition, existing topic discovery method sare often interfered by background words when mining sub-topics of events, which results in sub-topics of poor quality. And, existing method sare often unsupervised and thus it is difficult for them to ensure the intelligibility of the generated tag. For these difficulties, this paper proposes the sub-topic discovery method for omen events and the strategies for generating sub-topic tags. This paper proposes an ET-TAG model, which uses PLSA-BLM to discover sub-topic keywords, KL divergence to merge similar sub-topics, and then utilizes co-occurrence relations to update sub-topic keywords. Based on the sub-topic keywords, the external knowledge base is used to generate the corresponding labels for each sub-topic. Experiments on Sogou news corpus and specific multi-category public opinion events corpus can prove that ET-TAG has obvious advantages compared with traditional methods in sub-topic discovery. ET-TAG has higher accuracy when generating sub-topic tags.

This work is supported by the National Natural Science Foundation of China (61572473, 61472400), and the National Science Foundation for Young Scientists of China (61303156).