一种基于特征演变的新闻话题演化挖掘方法

赵旭剑[1] 杨春明[1] 李 波[1] 张 晖[1] 金培权[2] 岳丽华[2] 戴文锴[3]

[1](西南科技大学计算机科学与技术学院 四川 绵阳 621010)

[2](中国科学技术大学计算机科学与技术学院 合肥 230026)

[3](萨尔大学计算机科学系 萨尔布吕肯 德国 D-66123)

摘 要 话题演化挖掘研究可以准确完整地获取新闻话题动态演化各个阶段的话题内容,帮助用户理解新闻话题的来龙去脉以及话题内容之间的相关性和差异性,因此在网络新闻检索、网络舆情监控、互联网突发事件检测与应急管理等方面具有十分重要的作用和应用前景.现有工作由于缺乏对话题特征随时间发展而动态演变的深入分析,仅仅采用均值泛化的思想去增量扩充演化中的话题特征,引入大量话题无关信息,影响了话题关联的准确率,从而导致最终话题演化挖掘结果的偏斜.因此,针对以上问题,文中通过引入话题特征演变特性,提出一种针对话题演化的特征计算模型,在此基础上利用已有话题相关文档和最新文档进行话题信息动态增量扩充,通过对话题特征进行正向融合以及逆向过滤完成对特征信息的抗噪处理,提高话题关联的正确率,有效地解决了话题演化的偏斜问题.

关键词 话题演化;话题模型;演变特征;演化偏斜;社会计算;社交网络

中图法分类号 TP391 **DOI 号** 10.3724/SP.J.1016.2014.00819

A Topic Evolution Mining Algorithm of News Text Based on Feature Evolving

ZHAO Xu-Jian[1] YANG Chun-Ming[1] LI Bo[1] ZHANG Hui[1] JIN Pei-Quan[2]
YUE Li-Hua[2] DAI Wen-Kai[3]

[1](*School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, Sichuan* 621010)

[2](*School of Computer Science and Technology, University of Science and Technology of China, Hefei* 230026)

[3](*Department of Computer Science, Saarland University, D-66123 Saarbrücken, Germany*)

Abstract The research on the topic evolution mining can obtain the topic information accurately and completely at all topic episodes, which is able to help users understand the cause and effect as well as the correlation and difference of news topic. Thus, it has a very important role in Web News Search, Network Public Opinion Monitoring, Internet Incident Detection and Emergency Management, etc. Owning to lacking the in-depth analysis of the dynamic evolution of topic features over time in the existing work which only uses the mean generalization thought to extend topic features in the evolution process incrementally, a large number of irrelevant topic information is introduced into the current work. Meanwhile the low accuracy of the topic associated computation produced by current work leads to the deviation phenomena of the topic evolution mining. Aiming to deal with this issue, this paper first proposes a feature computation model of the topic-evolution-oriented through introducing the evolution characteristics of the topic feature, and then on this

收稿日期:2013-06-20;最终修改稿收到日期:2014-01-26.本课题得到国家自然科学基金(61202044,71273010)、四川省教育厅科研基金(12ZB326)、绵阳市网络融合工程实验室开放课题(12ZXWK04)、西南科技大学博士研究基金(12zx7116)资助.**赵旭剑**,男,1984 年生,博士,讲师,主要研究方向为中文信息处理、Web 信息检索.E-mail:Jasonzhaoxj@gmail.com.**杨春明**,男,1980 年生,硕士,讲师,主要研究方向为文本挖掘、知识工程.**李 波**,男,1977 年生,博士研究生,讲师,主要研究方向为信息过滤、信息安全.**张 晖**,男,1972 年生,博士,教授,主要研究领域为文本挖掘、知识工程.**金培权**,男,1975 年生,博士,副教授,主要研究方向为 Web 检索、知识管理.**岳丽华**,女,1952 年生,硕士,教授,主要研究领域为数据库、知识管理.**戴文锴**,男,1982 年生,博士研究生,主要研究方向为文本挖掘、机器学习.

basis, the article conducts the forward fusion and reverse filter under the existing topic-related stories and newly arrived stories in order to fulfill the incremental expansion of topic information and anti-noise processing. The experiment results show that this method improves the topic association precision and solve the topic evolution deviation problem effectively.

Keywords topic evolution; topic model; evolution feature; evolution deviation; social computing; social network

1 引 言

新闻作为一种流数据具有明显的动态变化性，而这种变化的载体就是新闻话题. 话题随着时间的发展而演化，反映了新闻事态阶段性渐变的过程. 而从认知心理学的角度来看，这样的演化过程正好体现了人类认知事物的一般逻辑顺序，当用户关注某个新闻话题的时候，他总是希望能够从了解新闻话题事件的缘由开始，逐步深入到事件的发展、曲折、高潮，最终到话题事件的结束. 整个逻辑顺序其实就是新闻话题完整的动态演化，是一种随时间渐变的过程. 因此，如果能够准确、完整地获取新闻话题在各个阶段的特征信息，并以话题事件时间为序，将各阶段话题内容全面整合，完成对新闻话题的动态演化挖掘，必然可以使人们更好地了解新闻话题(事件)的来龙去脉，不用阅读大量冗余的噪声新闻即可完全掌握新闻事态的发展.

同时，关注新闻话题的动态演化对于人们预警社会重大事故灾难也有着积极作用. 例如，图 1 反映了从 2002 年 11 月至 2003 年 3 月有关 SARS 疫情的新闻报道中我们抽取出的前 5 个关键词的变化趋势不难看出，随着时间的推移，疫情越发严重，就医人数从“两名”到“多例”，病情状况从“肺炎”到“重症”，而涉及的地区也更加广泛，从“佛山、河源”遍及到“广东省”，再从“广东”侵入到“广西、香港”. 这些话题特征随时间的变化充分体现了疫情的蔓延过程，准确地反映了事态的发展，因此，如果国家相关部门能够通过挖掘话题演化过程中新闻事件在内容和强度上的差异，重视这种变化对社会和人类带来的影响，我们就完全有可能在事件全面爆发之前的更早阶段采取应急措施，从而避免其产生更大的破坏和恶劣影响.

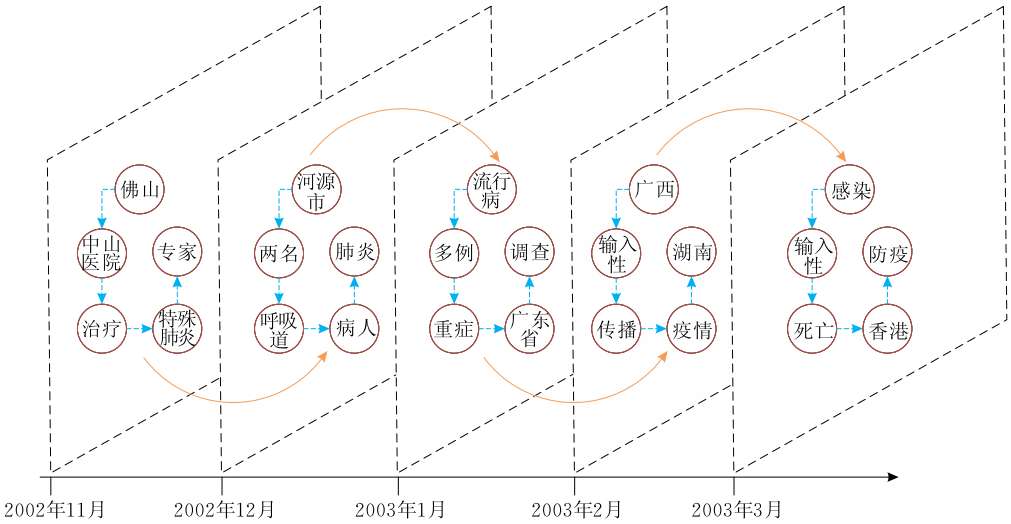


图 1 SARS 疫情相关新闻报道话题特征变化趋势

从实际应用来看，目前国内外各新闻门户网站①和新闻搜索引擎②几乎都提供了传统的新闻检索[1-2]、新闻分类[3-4]和新闻热度展示[5-6]等功能，但它们都局限在新闻话题的某一个状态，缺乏对新闻话题整个动态演化过程的深度理解，同时也忽视了新闻话题演变过程中各个状态之间的差异和关系对新闻话题检测与追踪的重要影响.

而对于传统的话题检测与追踪方法[7-12]，它们几乎都是以一个“点”的角度去观察新闻话题，只能

① http://news.people.com.cn; http://news.163.com; http://news.sina.com.cn; http://news.sohu.com

② http://news.baidu.com; http://news.google.com.hk

将新闻报道流中与某个话题相关的新闻聚类在一起，不能合理、有序地组织文档，更无法提供完整的新闻话题动态演化轨迹使用户直观、便捷、清晰地掌握新闻事件的因果关联和来龙去脉.

基于这样的实际需求，学术界近几年开始出现有关话题动态演化研究的工作[13-18]，并日益得到研究学者的关注. 根据文献分析，我们发现目前有关话题动态演化研究的工作主要集中于两类方法：一种是基于传统向量空间话题模型的话题演化挖掘[19-25]，通过将文档的时间信息作为话题属性引入到话题特征计算进而构建具有动态演变性的话题模型；另一种方法则是在概率话题模型的基础上，通过计算时间信息与话题、文档、词项的后验概率分布完成对话题演话在强度和内容上的追踪[14-15,17,26-29]. 但是，目前已有工作在话题模型动态更新问题上具有明显不足，它们未深入考虑已有词项特征与新增词项特征的演变特性对话题演化挖掘的影响，仅仅利用已有相关文档或者最新文档进行话题信息泛化扩充，未对话题模型进行抗噪处理，忽视了最新文档中噪声信息的引入和已有相关文档中噪声特征的放大，导致话题模型性能下降，影响最终话题演化挖掘的准确率. 图 2 描述了相关工作[23]中所得到的一个由于引入新文档中的噪声信息产生话题错误关联，导致话题演化偏斜的实例.

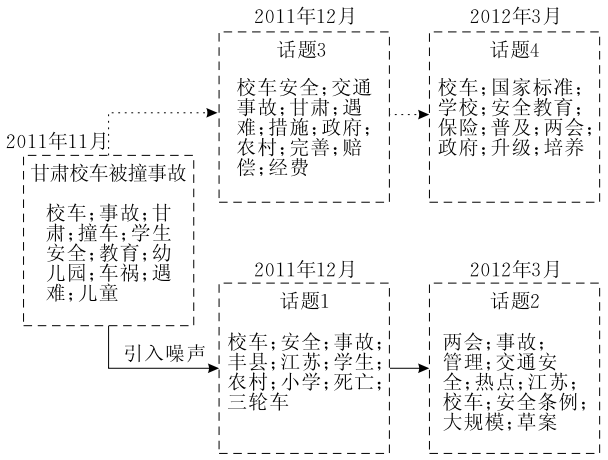


图 2　话题关联错误导致话题演化偏斜实例

该实例中由于话题“甘肃校车被撞事故”与之后的话题 1(“江苏丰县校车事故”)具有较强的语义关联性与特征一致性，因而算法错误地将两个话题关联起来并形成演化关系，影响了最后的话题演化挖掘结果.

因此，针对以上问题，本文我们提出了一种基于特征演变的话题演化挖掘方法，利用已有话题相关文档以及最新文档动态扩充话题特征，并采用正向融合和逆向过滤的思想进行模型抗噪处理，最后通过聚类完成新闻话题的演化轨迹挖掘. 与已有工作相比，我们的主要贡献在于：

(1) 提出了基于话题特征演变的特征计算理论模型，通过引入词项的突发性、连续性以及密集性，大大改善了话题模型的关联计算性能，显著提高了挖掘算法对于话题无关文档的判别以及话题内容阶段演变的识别.

(2) 在计算模型的基础上，给出了话题演化研究中基于已有话题相关文档和最新文档的话题模型增量式扩充策略，有效抑制了新噪声的引入和已有噪声的放大，提高了话题模型在动态演变过程中的抗噪性.

(3) 提出了面向单一话题新闻流和混合话题新闻流的话题演化挖掘机制，为真实环境下面向多源新闻的话题演化挖掘提供了统一的研究框架和方法.

2　话题演化特征分析

对于话题演化过程，无论是话题内容还是话题强度的动态变化性都是基于话题特征随时间演变而体现的. 同时，话题特征作为话题模型的构造元素，在很大程度上决定了话题关联计算的性能，直接影响了最终话题动态演化挖掘的结果. 因此，话题特征随话题演变而产生的变化差异实质就是话题演化的本质，充分研究话题特征的动态变化将有助于我们深入挖掘话题演化轨迹.

2.1　话题模型特征选择

目前针对文本话题的研究一般采用文档中的词项作为话题特征实体，而对于不同的研究对象，特征的选择以及权重计算也会产生变化. 针对网络新闻报道，我们在前期已经给出了针对网络新闻标题的话题特征抽取方法[30]，通过分析中文词性信息以及标题中的词素位置信息进行新词词典动态构建，同时结合新闻元素语言特征和标题中的词项结构完成话题词的抽取与话题权重计算，实验结果表明该方法较对比方法具有更高的准确率，抽取结果能较好地表示新闻话题内容. 然而，对于话题关联任务来说，需要对不同话题模型进行相似度计算，特征的全面性决定了话题信息的完整性，过少的话题特征将影响话题计算的性能[21]，因此，针对话题演化挖掘，我们增加了对新闻正文内容的考虑，全面扩充话题信息.

同时，为了在扩充话题特征的同时尽量减少噪声

信息的引入,我们对话题特征的来源进行了筛选.具体来说,我们将话题信息来源分为如表 1 所示的 9 类,然后基于传统的 TF×IDF 模型构建话题模型,在 500 篇中文新闻①组成的数据集上进行了 K-means 话题聚类实验,得到了如图 3 所示的实验结果.

表 1 话题信息来源分类

类别	话题信息来源
1	新闻标题(包括一级标题以及一级以下标题),下同
2	新闻正文第一句
3	新闻正文首段
4	新闻正文前两段
5	新闻正文全文
6	新闻标题+新闻正文第一句
7	新闻标题+新闻正文首段
8	新闻标题+新闻正文前两段
9	新闻标题+新闻正文全文

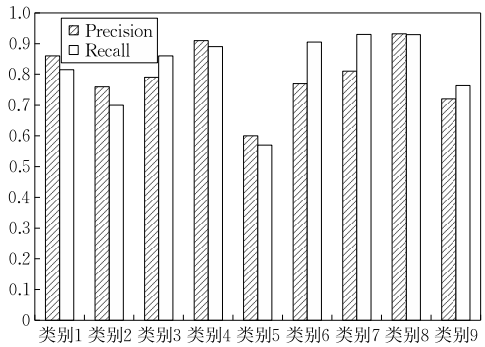


图 3 话题聚类实验结果

不难看出,由于新闻标题对于新闻话题的浓缩式体现,以及新闻报道的文学体裁所具有的篇章结构的特殊性(即新闻事件描述常常出现在新闻正文前两段),两者共同决定了基于新闻标题以及新闻正文前两段进行话题特征提取并构建话题模型的策略对于话题关联计算具有最优的性能表现.进一步分析实验结果,我们发现相比于标题与正文前两段,基于新闻全文的方法由于正文中非导语部分[31](通常为前两段之后的内容)可能包含多个子话题或者出现涉及话题某方面信息描述的内容,导致话题漂移,产生关联噪声,进而影响了话题聚类结果.因此,我们将采用基于新闻标题以及新闻正文前两段的选择策略进行话题演化挖掘的特征筛选,充分保证话题信息的完整性与准确性.

2.2 话题特征演变

话题演化过程中话题特征随着时间的发展产生动态的变化,不同阶段的话题特征所具有的差异性加大了传统话题关联方法对于噪声信息的检测难度,导致关联计算准确率的下降,进而使话题演化结果产生偏移.然而,从完整的话题演化轨迹来看,我们发现话题特征的动态变化具有一些特殊的演变特性,对于我们深入挖掘新闻话题演化特征具有重要的理论指导.

(1)突发性(Burst).不同阶段的话题内容在新闻话题演化过程中存在差异,具体表现为文档中词项的变化.在同一话题的新闻报道流中,新的词项出现往往意味着一个新的话题演化阶段的到来.描述同一个阶段话题的词项特征与其它阶段的话题特征具有明显差别,当一个阶段的主要话题特征发生改变时,新闻话题也发生了一次阶段演变.这对于在话题事件时间检测出现偏差的情况下判别两个具有相同时间的话题相关新闻是否存在演化关系具有指导意义.同时,强化词项特征的突发性权重,对于提高话题无关新闻报道的检测也具有显著作用.例如图 2 中讨论的"甘肃校车被撞事故",正是由于没有充分利用新词(如"丰县"、"江苏")在新闻报道中的突发特性对话题关联检测的区分功能,影响了已有话题模型对新噪声信息的过滤,导致关联计算的失误.

(2)连续性(Consecution).属于同一个话题演化阶段的新闻报道流由于具有相似的话题信息,通常含有一些相同的词项特征,我们把这些话题特征称为该阶段话题的轴心特征(Axis Features).轴心特征不会随着时间的发展而改变,而是始终贯穿于该演化阶段包含的所有报道.例如,一个关于"四川汶川地震"的新闻流 S,由 7 篇报道(d_1,d_2,d_3,d_4,d_5,d_6,d_7)组成,其中,连续出现在 d_2、d_3、d_4、d_5的词项"重建"和"援助"显然比仅仅出现在 d_1 和 d_4的"伤亡"和"治疗"具有更为显著的话题阶段信息标识特征,更能体现"灾后重建"这个演化阶段的话题含义,对于确定文档 d_2、d_3、d_4以及 d_5来源于同一个话题演化阶段具有十分重要的判别作用.因此,对词项进行连续性特征计算,可以帮助我们检测词项的阶段信息特性,有助于判断新闻流中不同文档的阶段归属一致性,同时,通过加强轴心特征在话题模型中的权重,可以达到相对弱化噪声特征的目的,从而提高话题演化融合的准确率.

(3)密集性(Intensity).来源于同一个演化阶段的文档除了具有相同的轴心特征以外,部分与话题关联度较高的词语在新闻报道中会高频出现.因而从整个演化过程来看,一些词语会集中地出现在某些特定时间范围内,使文档的词项特征在时间轴上

① 人民日报新闻语料(1998 年 1 月),北京大学计算语言研究所与富士通研究开发中心有限公司联合提供.

呈现出不同的分布. 例如，图 4 展示了话题“甘肃校车被撞事故”中 5 个词项特征在话题演变过程中的分布情况，由于它们都归属于“灾后救治，政府援助”这个演化阶段，因而在相应的时间范围内(2011 年 12 月)表现出较为一致的密集性. 因此，如果两篇新闻报道同时包含多个在特定时间区间内高频出现的词语，则它们很有可能归属于同一个话题演化阶段，这种特性将有助于我们对演化阶段信息的区分. 同时，对于话题无关文档，词项密集性在话题关联计算中也具有很强的话题关联检测能力，高频出现在特定时间区间内的词项往往标志着它所归属的文档与其它文档的话题差异.

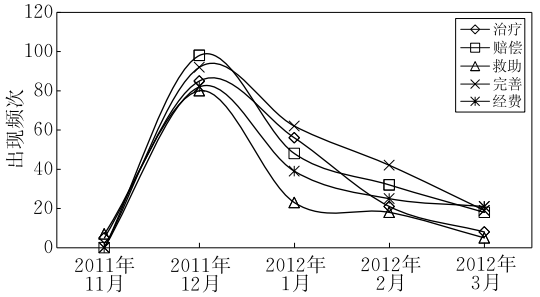


图 4　话题“甘肃校车被撞事故”特征词项分布

3　新闻话题动态演化挖掘

针对处理对象的不同，新闻话题动态演化挖掘可以分为两种模式，即针对同一话题新闻报道流的简单挖掘模式和针对混合话题新闻报道流的复杂挖掘模式. 换句话说，简单模式不需要话题无关检测，只需要对新闻报道流进行关联聚类；而复杂模式则需要通过话题关联计算区分话题无关文档，同时对话题进行相关文档追踪，再完成同一话题内新闻报道的关联融合. 图 5 给出了两种挖掘模式的完整流程图，复杂模式相对简单模式增加了对话题无关文档的检测，即对最新文档的分类操作，判断其是否归属已有话题.

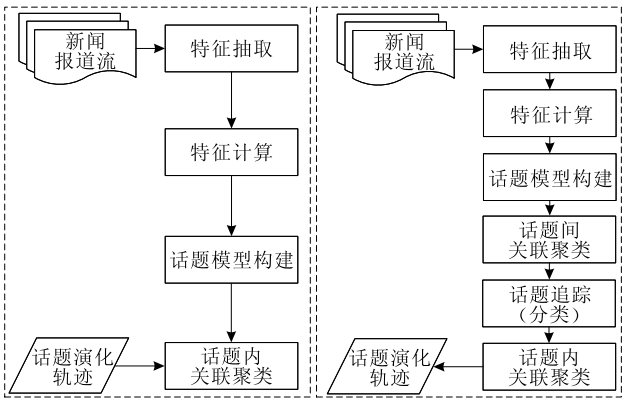


图 5　话题挖掘模式流程图

3.1　特征计算模型

由 2.2 节的讨论我们知道话题特征在话题动态演化过程中具有特殊的变化特性，这些特性充分体现了话题阶段性信息的差异，有助于识别话题演化阶段，同时，对于话题相关文档以及非相关文档的关联检测也具有十分重要的判别作用. 因此，我们根据话题特征的演变特性，提出了针对话题动态演化研究的特征计算模型，见式(1).

$$W(\|f\|)=W_O(f)+W_B(f)+W_C(f)+W_I(f) \quad (1)$$

特征 f 的权重 $W(\|f\|)$ 由 4 部分组成，分别是特征的原始权重 $W_O(f)$、突发权重 $W_B(f)$、连续权重 $W_C(f)$ 以及密集权重 $W_I(f)$. 其中，原始权重结合了前期工作中对新闻标题的特征计算模型[30]和传统的增量式 TF×IDF 模型[32]来表示词项在新闻文本中的基本权重，具体计算见式(2).

$$W_O(f)=\begin{cases} W_{itf\times idf}(f), & f \text{ 不在标题中} \\ W_{itf\times idf}(f)\times(1+W_{\text{title}}(f)), & f \text{ 在标题中} \end{cases} \quad (2)$$

公式中由于引入了词项基于标题的话题权重 $W_{\text{title}}(f)$，因而更适应于针对新闻文档的特征计算. 同时，对于词项特征在话题演化轨迹中的突发性，我们从词项与时间的独立性来进行分析，通过构建两者的列联表[3](如表 2)来计算 χ^2 独立性检验. 这里 N^{wt} 和 $N^{w\bar{t}}$ 分别表示在时刻 t 出现且包含词项 w 的新闻报道数以及时刻 t 之前出现的包含词项 w 的报道数，而 $N^{\bar{w}t}$ 和 $N^{\bar{w}\bar{t}}$ 则分别表示在时刻 t 出现且不含有词项 w 的报道数以及时刻 t 之前出现且不含有 w 的报道数.

表 2　词项 w 与时间 t 的列联表

	t	$\bar{t}$
w	N^{wt}	$N^{w\bar{t}}$
$\bar{w}$	$N^{\bar{w}t}$	$N^{\bar{w}\bar{t}}$

因此，χ^2 统计量可由式(3)求得，它表示了词项与时间的关联度，值越大代表词项越依赖于时间，越能体现词项 w 随着时刻 t 到来而突发出现的状态. 我们将归一化的词项 w_i 的 χ^2 统计值作为该特征的突发权重，见式(4).

$$\chi^2=\frac{(N^{wt}+N^{\bar{w}t}+N^{w\bar{t}}+N^{\bar{w}\bar{t}})(N^{wt}N^{\bar{w}\bar{t}}-N^{w\bar{t}}N^{\bar{w}t})^2}{(N^{wt}+N^{w\bar{t}})(N^{\bar{w}t}+N^{\bar{w}\bar{t}})(N^{wt}+N^{\bar{w}t})(N^{w\bar{t}}+N^{\bar{w}\bar{t}})} \quad (3)$$

$$W_B(f_i)=\chi^2(w_i)\Big/\sum_{j=1}^{M}\chi^2(w_j) \tag{4}$$

对于话题特征在同一个话题演化阶段所变现出来的连续性,我们结合前期关于话题事件时间的工作[33]采用式(5)进行计算.公式中具体符号定义见表3.这里的话题事件时间我们利用话题时间解析器(Topic Time Parser,TTP)[33]进行提取,该解析器通过挖掘新闻话题-时间关系模型,构建话题-时间关系树(Topic-Time Relationship Tree,TTRT),在此基础上采用基于话题权重和无监督学习的话题时间抽取算法自动抽取话题时间并进行时态表达规范化处理.

$$W_C(f)=\frac{(t_{\max}-t_{\min})/2}{1+\sum_{i=1}^{|S|}\log(t(d_{i+1})-t(d_i)+1)}\cdot\frac{1}{N} \tag{5}$$

表3 式(5)符号定义

符号	定义
S	报道流中含有特征 f 的文档集合
$t(d_i)$	包含特征 f 的第 i 个文档的话题时间
$t_{\max}$	已有报道集合中最新文档的话题时间
$t_{\min}$	已有报道集合中最早文档的话题时间
N	报道流中所有文档的数目

显然,通过式(5),对于连续出现在多个前后邻接的新闻报道中的词项较离散出现的词项具有更高的权重,有助于我们在关联融合中有效识别话题演化阶段标识.而在实际计算过程中,对于具有区间形式的话题事件时间,我们采用如下规则进行计算:

对 $\forall t=(t^s,t^e)$ 有

$$t_{i+1}-t_i=\begin{cases}(t^s_{i+1}-t^s_i)+(t^e_{i+1}-t^e_i), & t^s_{i+1}\neq t^e_{i+1}\text{且}\ t^s_i\neq t^e_i\\ t^s_{i+1}-t^e_i, & \text{其他}\end{cases} \tag{6}$$

此外,针对最新文档流和已有文档集合我们利用式(7)来挖掘话题特征的局部时间密集性.

$$W_I(f)=\frac{\sum_{d\in D}C_T(f,d)/(C_{T-1}(f)+C_D(f))}{\sum_{d\in D}\sum_{f\in d}C_T(f,d)/(C_{T-1}(f)+C_D(f))} \tag{7}$$

这里我们将文档流切分为多个时间窗口,最新的文档流属于窗口 T,T 中所有文档集合用 D 表示,特征 f 在最新文档集合 D 上出现的总次数为 $C_D(f)$,而 f 在 T 之前出现的次数则用 $C_{T-1}(f)$ 表示,因而特征 f 在当前时间窗口 T 内出现的次数与截止当前时间出现总次数的比值反映了 f 在整个时间轴上的局部集中现象,对于每个时间窗口 $W_I(f)$ 进行增量更新,保证了话题特征相对全局演化轨迹的密集性.

通过依次求得 $W_O(f)$、$W_B(f)$、$W_C(f)$ 以及 $W_I(f)$ 后,利用式(1)可得到特征 f 最终的话题权重 $W(\|f\|)$.不难看出,通过该计算模型,词项特征在话题演化过程中充分利用已有文档和最新文档进行动态更新,保证了特征演变与话题演化的一致性.具体来说,通过正向挖掘词项特征的突发性以及密集性,强化最新文档中新特征的话题权重,以及逆向挖掘词项特征在已有文档中具有的连续性,突出话题轴心特征,两方面共同作用有效提高了模型的抗噪能力,为针对话题演化研究的话题特征计算提出了具有指导意义的理论模型.

3.2 话题关联融合

由于向量空间模型(VSM)具有的表示直观、计算简单等优点,使其在 TDT 领域得到广泛研究和应用,同时相对概率模型,VSM 对于话题内容的演化挖掘具有更好的支持,因而我们采用 VSM 模型来表示新闻话题(文档).经过预处理(分词、词性标注、命名实体识别、停用词过滤)后的每个词项通过上一小节的计算成为话题(文档)向量的一维特征,所有词项组成的 N 维向量就代表了话题(文档)在特征空间的数字表示.

从话题关联计算的角度来分析整个话题演化挖掘过程,不难看出,无论是针对同一话题新闻流的简单挖掘模式还是混合话题新闻流的复杂挖掘模式,基于话题相似度计算的实质是始终不变的.复杂模式相对简单模式需要首先进行新话题的检测,对最新文档进行已知话题相关性的判别,即通过计算已知话题模型与最新文档模型的相似度来判断最新文档是否归属已知话题,然后对已知话题的文档集合进行话题演化聚类,即简单挖掘,因而复杂挖掘是一个两阶段模式,详细描述如算法1所示.

算法1. 混合话题模式的话题演化挖掘算法.

输入:最初到达的新闻报道流

输出:话题演化序列

```
1. BEGIN
2. ----------阶段 1----------
3.    FOR each s_i in story_set DO
4.    //构建话题模型
5.    ConstTopicModel(s_i);
6.    //采用 K-means 将所有已达到的新闻报道聚类
7.    KmeansCluster(story_set);
8.    FOR each new coming story s_i' in time slot n DO
```

```
9.    //计算 s_i' 和目标话题 T 的相似度 θ
10.   θ←SimCompute(s_i', T);
11.   //判断 s_i' 是否属于话题 T
12.   IF θ>θ_threshold
13.     //根据最新到达的话题相关的新闻报道更新话题模型
14.     ModelUpdate (T, s_i');
15.     //将s_i' 添加到目标话题 T 的报道集合 T_story_set
16.     T_story_set←AddtoTStorySet(s_i');
17.   ELSE
18.     //将 s_i' 添加到待处理话题的报道集合
19.     AddtoTopicSet(s_i');
20.   END IF
21. ----------阶段 2----------
22.  //采用 HAC 层次聚类法对所有与目标话题相关的新闻报道进行聚类，生成话题演化序列 S_Sequence
23.  S_Sequence←HACCluster(T_story_set);
24.  RETURN S_Sequence;
25. END
```

显然，算法中第 2 阶段的工作就是简单挖掘的整个过程，考虑到为了使聚类结果更好地体现话题演化的逻辑顺序，我们采用依照先后邻接合并规则(Successively Merge)聚类的 HAC 聚类方法[34]进行话题演化融合. 而对于最新到达的话题相关文档，利用 *ModelUpdate* 函数对已有话题模型进行增量扩充，包括新特征的添加以及已有特征权重的更新. 同时，基于文档的 VSM 模型表示方式，我们利用经典的余弦相似度(Cosine Similarity)公式来计算两个话题(文档)的关联度.

话题演化聚类操作完成后，与目标话题相关的新闻报道序列被组织成一系列类簇，每个类簇代表一个话题演化阶段，而整个报道序列则全面体现了目标话题的动态演化轨迹.

4　实验结果与分析

4.1　实验数据

由于本文提出的话题演化挖掘算法实际完成了话题检测与话题融合两个任务，因而我们分别采用 LDC(Linguistic Data Consortium)提供的 TDT3 语料①以及机器爬虫从新闻门户网站爬取的新闻专题语料作为实验数据集进行算法验证. 其中，TDT3 语料包含了新华社、联合早报以及 VOA Mandarin 3 个新闻源从 1998 年 10 月至 1998 年 12 月发布的共 12341 篇中文新闻，内容覆盖时事、财经、体育等 12 个领域，归属于 34 个人工标注的新闻话题，每个话题保证至少含有 4 篇相关新闻报道，话题描述由种子事件(事件属性)、话题说明、关联解释规则以及相关文档等信息组成(如图 6 所示). 这里由于 VOA Mandarin 新闻集主要针对语音识别处理，未对文本内容进行相应的语言处理，因而我们实际采用 5153 篇新华社新闻和 3817 篇联合早报新闻作为话题检测算法的实验数据集.

30009. Anti-Doping Proposals

Seminal Event

WHAT: The International Olympic Committee adopts a package of drug sanctions, and announces the formation of an anti-doping agency.

WHERE: Lausanne, Switzerland

WHEN: 11/27/98

Topic Explication

The IOC and 35 international sports federations adopted an agreement which precisely defines doping and imposes a minimum two-year suspension for athletes using steroids or other major performance enhancing drugs. Athletes caught a second time would be banned for life. **On topic**: Stories discussing the motivation for the new policy; adoption of the policy; specific details of the package; reactions of athletes, officials, nations to the new policy. Stories discussing particular athletes' drug use are not on topic unless they specifically relate to the new policy.

Rule of Interpretation Rule 9: New Laws

Related Articles: APW19981127.0499, APW19981202.1283

More examples: Yes, Brief.

图 6　话题描述实例

同时，为了更真实地检验算法性能，我们从新浪②、搜狐③、腾讯④三大新闻门户网站分别爬取了“SARS 事件”、“四川汶川地震”、“甘肃校车被撞事故”以及“江苏丰县校车事故”4 个话题的专题新闻数据作为话题演化挖掘算法的测试语料，数据集具体分布如表 4 所示. 这里考虑到由于专题新闻是基于人工编辑和话题分类的，因而可以在不考虑噪声

① http://projects.ldc.upenn.edu/TDT3/
② http://news.sina.com.cn/zt/
③ http://index.news.sohu.com/zhuanti/news_index.php
④ http://news.qq.com/topic/feature.htm

数据(话题无关文档)的情况下专门针对同一话题新闻报道流进行话题演化融合算法的评估,同时也可以灵活地将多个专题新闻数据集混合后进行话题演化复杂模式的挖掘算法检验.此外,"甘肃校车被撞事故"与"江苏丰县校车事故"两者同属一个主题,具有较强的话题相似性,在时间维度也存在紧密关联,因而对于话题检测算法具有更大的挑战,更能体现真实环境下复杂挖掘算法的实际性能.

表 4 专题新闻数据集分布

属性	SARS 事件			汶川地震			甘肃校车事故			江苏校车事故		
	新浪	搜狐	腾讯	新浪	搜狐	腾讯	新浪	搜狐	腾讯	新浪	搜狐	腾讯
文档数	1622	2188	1923	4189	3167	3999	2829	3613	3427	3024	1632	3918
总和		5733			11355			9869			8574	

4.2 评测机制

对于话题检测算法,我们采用美国国家标准与技术研究院(National Institute of Standards and Technology,NIST)针对 TDT 任务发布的评测标准,即检测错误代价 C_{det}(Detection Error Cost)来评估算法性能.在实际计算过程中,人们往往使用归一化后的检测错误代价 $Norm(C_{\text{det}})$来进行不同系统的性能比较.

同时,不难看出,话题检测算法的性能与关联阈值 $\theta_{\text{determine}}$的选取也有着十分重要的关系,TDT 评价标准采用检测错误权衡曲线图(Detection Error Tradeoff,DET)来表示漏检率和误检率随关联阈值不同选取而产生的变化趋势,更直观地反映检测算法的整体性能,在随后的实验评测中我们将利用 DET 对实验结果进行科学分析.

此外,针对话题演化挖掘算法的效能评估,我们采用 Wei 等人[20]提出的评价标准进行实验评测.该标准将各个话题演化阶段中的报道成对按组合方式取出,在此基础上通过计算聚类的准确率 P_c、召回率 R_c以及 F_c评估算法性能,3 个指标具体定义如下所示.

$$P_c=\frac{\text{由算法产生的话题演化所形成的报道对且同时属于正确话题演化所形成的报道对的个数}}{\text{由算法产生的话题演化所形成的报道对组合总数}},$$

$$R_c=\frac{\text{由算法产生的话题演化所形成的报道对且同时属于正确话题演化所形成的报道对的个数}}{\text{正确话题演化所形成的报道对组合总数}},$$

$$F_c=\frac{2\times P_c\times R_c}{P_c+R_c}.$$

对于最终话题演化挖掘结果的评测,我们采用百度百科①与维基百科②相应话题的内容描述对算法实验结果进行人工评判.两个在线知识库都是基于人工编辑,按照一定的新闻事件发展顺序(如图 7 中维基百科对于"甘肃校车被撞事故"的话题描述目录)给出新闻话题多个层面的相关描述,因而能较好地反映话题演化过程,对于文本的话题演化挖掘算法具有科学的评测指导意义.

图 7 维基百科关于"甘肃校车事故"的话题描述

4.3 实验设计

针对本文的话题演化挖掘算法性能评测,我们将实验分为 3 个部分,分别是基于 TDT3 数据集和混合专题新闻语料的话题检测评估、基于单一专题新闻语料的话题融合(简单挖掘)评估以及基于混合专题新闻语料的话题演化挖掘(复杂挖掘)评估.

① http://baike.baidu.com/
② http://zh.wikipedia.org/

实验中我们分别实现了 6 种对比方法来进行本文算法的效能评测，它们分别是基于传统增量式 TF×IDF 计算模型的话题演化挖掘方法 $M_{itf\times idf}$[32]、在 $M_{itf\times idf}$ 基础上引入时间衰减因子[10,32,35]的 $M_{itf\times idf-decay}$、基于时间特征 TF ×IDF 计算模型的话题演化挖掘算法 $M_{tf\times idf-Tempo}$[20-21]、在 $M_{tf\times idf-Tempo}$ 基础上引入时间衰减因子的 $M_{tf\times idf-Tempo-decay}$、基于 LDA 话题模型的话题演化挖掘方法 $M_{LDA-assoc}$[26]以及以子话题关联来挖掘话题演化的 M_{ELDA}[27]. 具体实验布置见表 5.

表 5　专题新闻数据集分布

实验单元	数据集	参与对比方法
话题检测评估	TDT3＋混合专题新闻	$M_{itf\times idf}$，$M_{itf\times idf-decay}$，$M_{tf\times idf-Tempo}$，$M_{tf\times idf-Tempo-decay}$
话题融合评估	单一专题新闻	$M_{itf\times idf}$，$M_{itf\times idf-decay}$，$M_{tf\times idf-Tempo}$，$M_{tf\times idf-Tempo-decay}$，$M_{LDA-assoc}$，$M_{ELDA}$
话题演化挖掘评估	混合专题新闻	$M_{itf\times idf}$，$M_{itf\times idf-decay}$，$M_{tf\times idf-Tempo}$，$M_{tf\times idf-Tempo-decay}$，$M_{LDA-assoc}$，$M_{ELDA}$

这里我们采用 Nallapati 等人[35]提出的时间衰减函数来计算衰减因子，进而调整话题相似度的比较，具体计算见式(8).

$$sim_{decay}(d_i,d_j)=sim(d_i,d_j)\times\frac{1}{\exp(|t(d_i)-t(d_j)|/T)} \quad (8)$$

该公式反映了两个文档时间距离越大，其描述同一话题的可能性越小，这一理论在传统 TDT 研究中具有较好的性能表现，因而我们将其作为模型属性分别引入到相应算法中与本文算法进行对比.

此外，由于 $M_{LDA-assoc}$ 算法与 M_{ELDA} 算法主要针对话题演化研究，因而在话题检测评估单元中我们未对该方法进行比较实验.

4.4　实验评测

4.4.1　话题检测实验评测

首先，我们基于 TDT3 数据集对本文提出的基于特征演变的新闻话题演化挖掘方法 $M_{itf\times idf-feature}$ 以及 4 种对比算法进行了比较实验，得到图 8 所示的 DET 曲线图.

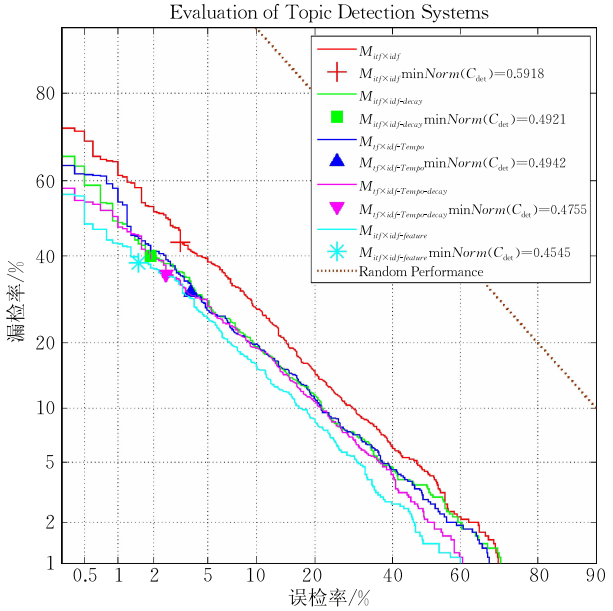


图 8　基于 TDT3 数据集的话题检测算法性能评测

由于检测错误代价由系统的漏检率(Miss Probability)和误检率(False Alarm Probability)两部分线性组合得到，因而 DET 曲线图中越靠近坐标空间左下方的曲线越具有较好的系统整体检测性能. 同时，曲线上最小的检测错误代价归一化结果 min$Norm(C_{det})$ 代表了对应系统的最优性能，也是评价话题检测方法的重要指标. 从图 8 可以看出，本文提出的 $M_{itf\times idf-feature}$ 方法较对比算法具有更优的整体性能，同时，最小检测错误代价在 4 种对比方法的基础上最多减少了 23.2%，最小也降低了 4.42%，充分说明了特征演变因素对于话题检测的有效性.

此外，我们将 4 种新闻专题语料按时间窗口进行划分，并随机依次从各个窗口内抽取新闻报道组成混合数据集进行算法实验，评测结果如图 9 所示.

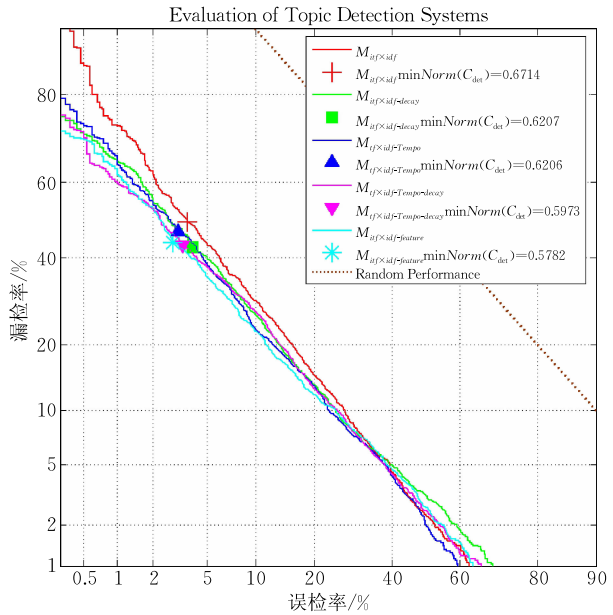


图 9　基于混合专题新闻数据集的话题检测算法性能评测

对比 TDT3 数据集的实验结果，我们发现测试算法在混合专题新闻数据集上的实验性能都有所下降，除了 $M_{itf\times idf}$ 方法外，其余 4 种方法的最小检测错误代价相对集中，差距很小，特别是引入时间连续

性特征的 $M_{tf\times idf-Tempo}$ 方法反而较未考虑轴心特征的 $M_{itf\times idf-decay}$ 方法在最优性能上略有下降. 分析原因,由于"甘肃校车事故"与"江苏校车事故"语料的时间跨度分别为(2011 年 11 月～2012 年 3 月)与(2011 年 12 月～2012 年 3 月),因而两部分语料具有紧密的时间连续性,导致"校车"、"车祸"等具有语义关联且同时出现在两个不同话题但时间连续的报道中的噪声词项具有较高权重,影响了算法性能. 因此,我们针对话题检测任务,将特征计算模型修改为原始权重、突发权重以及密集权重之和,暂不考虑连续权重,得到如图 10 所描述的对比实验结果.

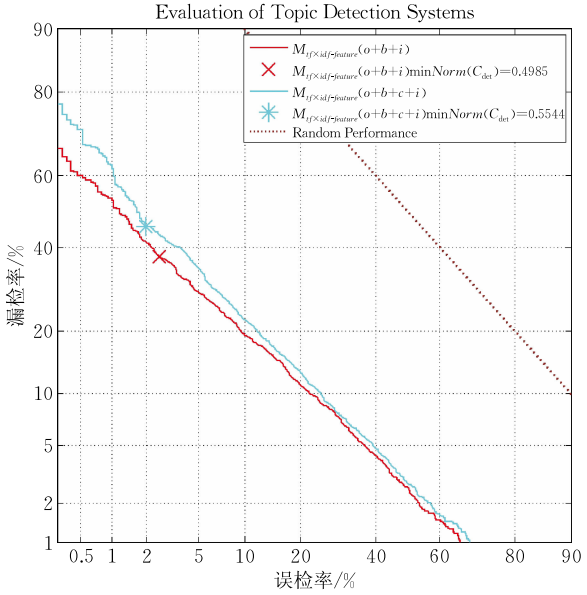


图 10 基于不同计算模型的话题检测算法性能评测

通过弱化出现在连续报道中的噪声特征,不考虑演变连续性的 $M_{itf\times idf-feature}(o+b+i)$ 方法较基于 3 种演变特性的 $M_{itf\times idf-feature}(o+b+c+i)$ 方法在最优性能上提高了 10.08%,同时算法在整个阈值空间也具有更为理想的整体系能,对面向话题检测的特征计算模型构建提供了一种新的思路.

4.4.2 话题融合实验评测

我们将 4 个专题新闻语料分别作为实验数据集对话题融合(简单挖掘)进行了算法性能评测. 其中,对于 $M_{LDA-assoc}$ 方法,我们采用了与原始文章一样的参数配置,即 $\alpha=50/T$,$\beta=0.1$,T 为话题数目,Gibbs 抽样迭代次数 $N=300$;而对于 M_{ELDA} 方法,由于我们不需要判别子话题演化的类型(产生、消亡、继承、分裂以及合并),因此只对检测出的子话题进行了关联结算,其中权重向量包含的每个元素取值为 0.5,相似度阈值为 -2. 具体实验结果如表 6 所示.

不难看出,相对 6 种对比方法,本文提出的基于特征演变的话题演化挖掘方法在单一话题数据集上具有最优的算法性能. 具体分析,由于我们在话题特征计算过程中充分考虑了特征的演变特性,因而较传统增量式 TF×IDF 模型更适合话题演化融合,而相对 $M_{tf\times idf-Tempo}$ 方法,我们除了引入已有文档的连续权重以外还融合了最新文档的突发权重和密集权重,因而具有更强的话题抗噪性,并且到达甚至超过了引入时间衰减因子的效果. 此外,$M_{LDA-assoc}$ 方法在整体性能上具有不稳定性,相对时间跨度较短且内

表 6 话题融合实验评测结果

算法	SARS 事件			汶川地震			甘肃校车事故			江苏校车事故		
	P_c	R_c	F_c	P_c	R_c	F_c	P_c	R_c	F_c	P_c	R_c	F_c
$M_{itf\times idf}$	0.64	0.63	0.63	0.61	0.65	0.63	0.58	0.62	0.60	0.62	0.64	0.63
$M_{itf\times idf-decay}$	0.69	0.76	0.72	0.64	0.64	0.64	0.58	0.62	0.60	0.62	0.62	0.62
$M_{tf\times idf-Tempo}$	0.66	0.67	0.66	0.66	0.67	0.66	0.69	0.72	0.70	0.75	0.72	0.73
$M_{tf\times idf-Tempo-decay}$	0.69	0.75	0.72	0.65	0.58	0.61	0.70	0.65	0.67	0.76	0.77	0.76
$M_{LDA-assoc}$	0.58	0.64	0.61	0.71	0.69	0.70	0.67	0.66	0.66	0.69	0.68	0.68
M_{ELDA}	0.62	0.64	0.63	0.69	0.69	0.69	0.65	0.62	0.63	0.64	0.63	0.63
$M_{itf\times idf-feature}$	0.73	0.75	0.74	0.71	0.69	0.70	0.70	0.72	0.71	0.77	0.76	0.76

容相关度较高的新闻流容易产生大量关联错误,导致最终算法性能下降. 而 M_{ELDA} 方法则由于机械式地采用新闻发布时间作为新闻话题时间,导致子话题内容与时间映射产生错误,进而影响了最终的挖掘结果.

考虑到上一小节特征计算模型中连续权重对于话题检测性能的影响,我们在本实验单元对两种方式的计算模型也进行了对比实验,准确率和召回率的比较结果如图 11 所示.

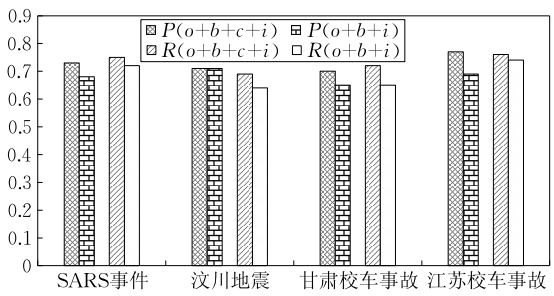


图 11 基于不同计算模型的话题融合算法性能比较

由于话题特征的连续权重在话题融合过程中对已有文档含有的轴心特征进行权重强化，有效抑制了已有噪声信息的放大，因此较仅仅考虑突发权重和密集权重的计算模型取得了更理想的实验结果，没有产生在话题检测中由于对时间连续但话题无关的文档关联错误而导致性能下降的问题.

4.4.3　话题演化挖掘实验评测

为了更有针对性地评测复杂演化挖掘方法在实际应用中的普适性，本实验单元我们将数据集划分为两组，即“SARS 事件”、“汶川地震”和“甘肃校车事故”组成的第 1 组以及“甘肃校车事故”和“江苏校车事故”组成的第 2 组. 显然，第 2 组数据由于同属于一个主题(领域)，存在紧密的话题语义关联，并且时间跨度较小，具有比较宽泛的时间重叠，因而相对第 1 组数据对算法性能更具挑战性.

首先表 7 给出了本文算法在第 1 组数据集上针对“汶川地震”的话题演化挖掘实验结果.

表 7　话题“汶川地震”演化挖掘实验结果

演化阶段标识	时间范围	报道数	前 10 个特征词
E_1	(2008.5，2008.5)	5677	汶川、地震、灾区、安排、四川、抢救、温家宝、里氏、5 月 12 日、解放军
E_2	(2008.6，2008.6)	3179	汶川、北川、对口支援、失踪人口、政府、物资、团结、过渡房、援助、民政部
E_3	(2008.7，2008.10)	1135	重建、经济援助、住房、学校、汶川、灾民、就业保障、损失、小学、安全教育
E_4	(2008.11，2008.11)	340	广东、基础设施、灾害、资金、永久性住房、解放军、稳定、上海、自然环境、全国
E_5	(2008.12，2009.2)	795	北川、汶川、抗震、春节、新房、对口援建、质量监督、心理辅导、政府、风景区
E_6	(2009.3，2009.4)	71	当地政府、地震遗址、旅游、财政、保护、返乡、就业、重建、四川、余震
E_7	(2009.5，2009.5)	158	对口援建、住房、城市建设、文化产业、医疗卫生、基础、管理、生活、悼念、失踪

从表格数据可以看出，算法最终挖掘结果具有一定的合理性并且与人类认知逻辑较为一致. 在地震发生后的两个月集中了整个新闻集大约 78% 的报道量，此时人们最关注的是灾害情况和抢救结果；而从 7 月至 11 月，随着时间推移，人们的关注度有所下降，该阶段报道最多的是灾后救助、对口支援等情况；随着 2009 年春节的到来，又有不少媒体开始关注该话题，报道灾后重建的进展；由于地震一周年的原因，2009 年 5 月较 3 月和 4 月有较多报道量，“悼念”、“失踪”等词项反映了该时段话题演化阶段的特征.

此外，针对第 2 组数据集，我们将本文算法的实验结果与对比方法进行了比较，具体评测结果如表 8 所示.

表 8　“甘肃校车事故”与“江苏校车事故”混合话题演化挖掘评测结果

算法	甘肃校车事故			江苏校车事故		
	P_c	R_c	F_c	P_c	R_c	F_c
$M_{itf\times idf}$	0.52	0.51	0.5150	0.56	0.52	0.5393
$M_{itf\times idf-decay}$	0.53	0.55	0.5398	0.58	0.57	0.5750
$M_{tf\times idf-Tempo}$	0.52	0.59	0.5528	0.62	0.59	0.6046
$M_{tf\times idf-Tempo-decay}$	0.58	0.56	0.5698	0.62	0.59	0.6046
$M_{LDA-assoc}$	0.42	0.51	0.4606	0.63	0.61	0.6198
M_{ELDA}	0.49	0.52	0.5046	0.58	0.62	0.5993
$M_{itf\times idf-feature}$	**0.64**	**0.62**	**0.6298**	**0.65**	**0.62**	**0.6346**

这里的挖掘方法在第一步的话题检测任务中我们都采用了 4.4.1 节中讨论的新的计算模型进行特征权重计算，而在话题融合过程中则还是采用原始的计算模型. 将表 8 与表 6 的数据进行比较不难看出，由于话题无关文档的影响，复杂挖掘模式下的算法性能较简单模式有明显下降，但本文提出的基于特征演变的挖掘方法还是取得了最佳评测结果. 词项突发性和密集性的引入，使话题模型具有更精确的话题动态描述，相对仅仅考虑已有文档中时间与特征局部关系的 $M_{tf\times idf-Tempo}$ 方法和 $M_{tf\times idf-Tempo-decay}$ 方法，无论是在话题检测还是话题融合两个阶段都表现出更好的关联计算准确性.

同时，轴心词项在话题演化过程中呈现出的连续性使我们的挖掘算法较基于增量式 TF×IDF 模型的 $M_{itf\times idf}$ 方法和 $M_{itf\times idf-decay}$ 方法在针对已有话题模型的特征增量扩充过程中具有更强的噪声弱化处理能力，进一步提高了挖掘算法在话题融合阶段的系统性能.

此外，“甘肃校车事故”与“江苏校车事故”在时间与语义上的紧密关联使 $M_{LDA-assoc}$ 方法与 M_{ELDA} 方法产生较明显的话题或者子话题关联计算错误，导致严重话题偏斜(如图 2 所示)，使算法准确性急剧下降. 同时，由于 LDA 模型未考虑词项对话题表示的有效性与合法性，导致算法产生一些无意义的“垃圾”话题，例如由“责任、生命、监督、调查、发生、标准、计划、安全、研究、法制”等词项组成的无意义话题，它们使包含其中词项的文档错误关联，影响了算法整体性能. 此外，无论是 $M_{LDA-assoc}$ 方法与 M_{ELDA} 方法或者其它的基于概率主题模型的方法，都是以词

语为话题单位,但它们并没有考虑词语的话题角色与语义,将词语仅仅作为一个表征(Token)引入话题模型的计算,忽视了话题的完整性.而本文算法由于是基于新闻元素对象进行特征抽取,因而充分保证了话题特征的完整性和语义性,有效弥补了 $M_{LDA-assoc}$ 方法与 M_{ELDA} 方法的不足.

5 结束语

针对已有话题演化挖掘方法在模型特征计算与模型动态更新上的不足,本文我们提出了一种基于特征演变的新闻话题演化挖掘方法,通过引入词项特征在话题演化过程中的变化特性,构建增量式特征计算模型,并且利用已有话题相关文档和最新文档进行话题特征的正向融合和逆向过滤,显著提高了话题模型的准确率,充分改善了关联计算的整体性能,进而有效提高了最终挖掘结果的正确性和完整性.

下一步我们的工作将围绕话题演化复杂挖掘的自适应以及模型优化两方面展开,探讨话题演化在更多应用背景下的系统研究框架和方法.

致 谢 审稿老师给出了宝贵的评语和修改意见,在此表示感谢!

参 考 文 献

[1] Leonidas K, Vassilis K, Isambo K. Semantic search in the world news domain using automatically extracted metadata files. Journal of Knowledge-Based Systems, 2012, 27: 38-50

[2] Ronald P. Facing scalability: Naming faces in an online social network. Pattern Recognition, 2012, 45(6): 2335-2347

[3] Wang C, Zhang M, Ma S, et al. Automatic online news issue construction in Web environment//Proceedings of the WWW. Beijing, China, 2008: 457-466

[4] Wang C, Zhang M, Ru L, et al. Automatic online news topic ranking using media focus and user attention based on aging theory//Proceedings of the CIKM. Napa Valley, USA, 2008: 1033-1042

[5] Cui W, Liu S, Tan L, et al. TextFlow: Towards better understanding of evolving topics in text. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2412-2421

[6] Kim D, Oh A. Topic chains for understanding a news corpus//Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing). Tokyo, Japan, 2011: 163-176

[7] Lavrenko V, Allan J, DeGuzman E, et al. Relevance models for topic detection and tracking//Proceedings of the Human Language Technology Conference (HLT). San Diego, USA, 2002: 104-110

[8] Yang Y, Ault T, Pierce T, et al. Improving text categorization methods for event tracking//Proceedings of the ACM SIGIR'00. Athens, Greece, 2000: 65-72

[9] Yang Y, Carbonell J, Brown R, et al. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval, 1999, 14(4): 32-43

[10] Allan J, Carbonell J, Doddington G, et al. Topic detection and tracking pilot study: Final report//Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. Virginia, USA, 1998: 194-218

[11] Hong Yu, Zhang Yu, Fan Ji-Li, et al. New event detection based on division comparison of subtopic. Chinese Journal of Computers, 2008, 31(4): 687-695(in Chinese)
(洪宇, 张宇, 范基礼等. 基于子话题分治匹配的新事件检测. 计算机学报, 2008, 31(4): 687-695)

[12] Zhang Xiao-Ming, Li Zhou-Jun, Chao Wen-Han. Research of automatic topic detection based on incremental clustering. Journal of Software, 2012, 23(6): 1578-1587(in Chinese)
(张小明, 李舟军, 巢文涵. 基于增量型聚类的自动话题检测研究. 软件学报, 2012, 23(6): 1578-1587)

[13] Yu Man-Quan, Luo Wei-Hua, Xu Hong-Bo, et al. Research on hierarchical topic detection in topic detection and tracking. Journal of Computer Research and Development, 2006, 43(3): 489-495(in Chinese)
(于满泉, 骆卫华, 许洪波等. 话题识别与跟踪中的层次化话题识别技术研究. 计算机研究与发展, 2006, 43(3): 489-495)

[14] Hall D, Jurafsky D, Manning C D. Studying the history of ideas using topic models//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Honolulu, Hawaii, USA, 2008: 363-371

[15] Blei D, Lafferty J. Dynamic topic models//Proceedings of the International Conference on Machine Learning (ICML). Pittsburgh, USA, 2006: 113-120

[16] Li B, Li W, Li Q. Enhancing topic tracking with temporal information//Proceedings of the ACM SIGIR. Seattle, Washington, USA, 2006: 667-668

[17] Wang X, McCallum A. Topic over time: A non-Markov continuous-time model of topical trends//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 424-433

[18] Giannella C, Han J, Pei J, et al. Mining frequent patterns in data streams at multiple time granularities//Proceedings of the NSF Workshop on Next Generation Data Mining. Toronto, Canada, 2003: 191-212

[19] Zhao Hua, Zhao Tie-Jun, Yu Hao, et al. Dynamic evolvement-oriented topic detection research. High Technology Letters, 2006, 16(12): 1230-1235(in Chinese)
(赵华，赵铁军，于浩等. 面向动态演化的话题检测研究. 高技术通讯，2006，16(12)：1230-1235)

[20] Wei C P, Lee Y H, Chiang Y S, et al. Discovering event episodes from news corpora: A temporal-based approach//Proceedings of the ACM ICEC'09. Taipei, China, 2009: 72-80

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

[22] Ma N, Yang Y, Rogati M. Applying CLIR techniques to event tracking//Proceedings of the Information Retrieval Technology, Asia Information Retrieval Symposium. Beijing, China, 2004: 24-35

[23] Larkey L S, Feng Fangfang, Connell M, et al. Language-specific models in multilingual topic tracking//Proceedings of the SIGIR 2004. Sheffield, UK, 2004: 402-409

[24] Lo Y Y, Gauvain J L. The LIMSI topic tracking system for TDT2001//Proceedings of the Topic Detection and Tracking Workshop 2001. Gaithersburg, Maryland, USA, 2001: 1-5

[25] Hong Yu, Cang Yu, Yao Jian-Min, et al. Descending kernel track of static and dynamic topic models in topic tracking. Journal of Software, 2012, 23(5): 1100-1119(in Chinese)
(洪宇，仓玉，姚建民等. 话题跟踪中静态和动态话题模型的核捕捉衰减. 软件学报，2012，23(5)：1100-1119)

[26] Chu Ke-Ming, Li Fang. Topic evolution based on LDA and topic association. Journal of Shanghai Jiaotong University, 2010, 44(11): 1496-1500(in Chinese)
(楚克明，李芳. 基于LDA话题关联的话题演化. 上海交通大学学报，2010，44(11)：1496-1500)

[27] Hu Yan-Li, Bai Liang, Zhang Wei-Ming. Modeling and analyzing topic evolution. Acta Automatica Sinica, 2012, 38(10): 1690-1697(in Chinese)
(胡艳丽，白亮，张维明. 一种话题演化建模与分析方法. 自动化学报，2012，38(10)：1690-1697)

[28] Xu Ge, Wang Hou-Feng. The development of topic models in natural language processing. Chinese Journal of Computers, 2011, 34(8): 1423-1436(in Chinese)
(徐戈，王厚峰. 自然语言处理中主题模型的发展. 计算机学报，2011，34(8)：1423-1436)

[29] Cao Juan, Zhang Yong-Dong, Li Jin-Tao, et al. A method of adaptively selecting best LDA model based on density. Chinese Journal of Computers, 2008, 31(10): 1780-1787(in Chinese)
(曹娟，张勇东，李锦涛等. 一种基于密度的自适应最优LDA模型选择方法. 计算机学报，2008，31(10)：1780-1787)

[30] Zhao X, Jin P, Yue L. A novel POS-based approach to Chinese news topic extraction from Internet//Proceedings of the Conference on Database Theory and Application. Sanya, China, 2008: 39-42

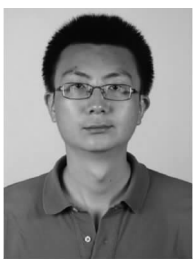
[31] Zhong Zhi-Yuan. Internet Journalism. Beijing: Peking University Press, 2002(in Chinese)
(仲志远. 网络新闻学. 北京：北京大学出版社，2002)

[32] Yang Y, Pierce T, Carbonell J. A study on retrospective and on-line event detection//Proceedings of the ACM SIGIR'98. Melbourne, Australia, 1998: 28-36

[33] Zhao Xu-Jian, Jin Pei-Quan, Yue Li-Hua. TTP: A topic time parser on Chinese news from Internet. Journal of Chinese Computer Systems, 2013, 34(5): 1042-1049(in Chinese)
(赵旭剑，金培权，岳丽华. TTP：一个面向中文新闻网页的主题时间解析器. 小型微型计算机系统，2013，34(5)：1042-1049)

[34] Manning C D, Raghavan P, Schtze H. Introduction to Information Retrieval. New York, USA: Cambridge University Press, 2008

[35] 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 (CIKM). Washington, USA, 2004: 446-453



ZHAO Xu-Jian, born in 1984, Ph.D., lecturer. His research interests include Chinese information processing, Web information retrieval.

YANG Chun-Ming, born in 1980, M. S., lecturer. His research interests include text mining, knowledge engineering.

LI Bo, born in 1977, Ph. D. candidate, lecturer. His research interests include information filtering, information security.

ZHANG Hui, born in 1972, Ph. D., professor. His research interests include text mining, knowledge engineering.

JIN Pei-Quan, born in 1975, Ph.D., associated professor. His research interests include Web retrieval, knowledge management.

YUE Li-Hua, born in 1952, M. S., professor. Her research interests include database, knowledge management.

DAI Wen-Kai, born in 1982, Ph. D. candidate. His research interests include text mining, machine learning.

Background

The research on topic evolution mining, as an advanced text mining technology, has a very important role in Web News Search, Network Public Opinion Monitoring, Internet Incident Detection and Emergency Management, etc. However, nowadays topic evolution mining is faced with a major challenge, the low accuracy of the topic associated computation, which leads to the deviation phenomena of the topic evolution mining.

This paper analyzes this challenge, and surveys the relevant research work that aimed to address the challenge. And then this paper proposes a topic evolution mining algorithm of news text based on feature evolving, which improves the topic association precision and solve the topic evolution deviation problem effectively.

This paper is supported in part by National Natural Science Foundation of China under Grant Nos. 61202044 and 71273010, by the Scientific Research Funds in Sichuan Province Department of Education under Grant No. 12ZB326, by the Open Projects Program of Engineering Lab of Network Convergence of Mianyang under Grant No. 12ZXWK04, by Doctoral Research Fund of SWUST under Grant No. 12zx7116.

Researchers started the research on topic evolution of Web texts in 2008, and have achieved some promising results involved in topic model, temporal expression normalization and topic evolution detection. More detailed can be found on the author's publications.