基于转发关系的微博话题演化算法

徐 伟 赵 斌 吉根林
(南京师范大学计算机科学与技术学院 南京 210023)

摘 要 现有的话题演化研究主要针对长文本。因此研究了微博短文本的话题演化问题,综合考虑微博的文本特征和非文本特征,利用微博的传播特性,提出了基于转发关系的微博话题演化算法MTERR。该算法首先以话题模型为基础,结合微博转发特性和话题的时间特征提取微博话题;然后采用话题的内容相关性指标和转发关联度指标构建话题关联函数,生成话题演化拓扑图;最后,基于真实微博数据集的实验结果表明,MTERR算法生成的话题演化图可以有效地反映热点事件发展演化的过程。

关键词 微博,话题演化,短文本,话题模型
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.2.017

Microblog Topic Evolution Algorithm Based on Retweeting Relationship

XU Wei ZHAO Bin JI Gen-lin
(School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023, China)

Abstract Existing work has been focused on topic evolution of long text. This paper aimed to that of short text. We proposed a microblog topic evolution algorithm MTERR based on retweeting relationship. Firstly, we utilized a topic model to obtain topic information from microblog messages by combining retweeting features and time characteristics. Then, we built a topic correlation function to generate a topic evolution topological graph by incorporating topic content and retweeting relationship. Experiments on the real-world microblog datasets show the feasibility and effectiveness of our proposed method.

Keywords Microblog, Topic evolution, Short text, Topic model

1 引言

话题演化研究是自然语言处理中的研究热点。传统的话题演化研究主要基于话题模型[1-3],将文字信息以向量形式表示,简化了话题模型求解的复杂度。然而由于没有考虑文字信息之间的语义关联,因此导致潜在的词项关联信息被忽略。近年来,为了弥补话题模型所带来的语义缺失,越来越多的话题演化研究开始采用概率主题模型。经典的概率主题模型包括概率隐性语义索引模型(probabilistic Latent Semantic Indexing, pLSI)[4]和潜在狄利克雷分布模型(Latent Dirichlet Allocation, LDA)[5]。

话题关联是话题演化研究中的一个重要问题。话题模型主要基于静态视角发现话题信息,而话题本身是一种数据流形式的文本抽象,它随着时间推移而不断演化。因而,在充分考虑时间因素的前提下,关联不同时间段的话题,可以有效重建话题演化过程。目前,话题关联方法分为两类:1)在模型中直接考虑话题之间的关联;2)通过构造关联函数实现话题关联。第一种关联方法是对话题模型进行改造,充分考虑时间信息,构建话题演化模型。根据时间信息引入方式的不同,话题演化模型又分为3种类型:后离散时间型(Time Post-discretized)[6]、先离散时间型(Time Pre-discretized)[7]和时间变量结合型(Time Variable Joint)[8]。后离散时间型首先针对整个文档集建立话题模型,然后再根据文档的时间属性,将文档划分到不同的时间窗口中,利用各个时间窗口内话题的强度变化来表示话题的演化。该模型步骤简单,计算代价小且对封闭型数据集有较好的性能表现,但是由于存在基于文档无序的假设,使得其忽略了话题本身的有序性。先离散时间型首先将文档按照一定的时间粒度划分到相应的时间窗口中,然后分别对各个时间窗口内的文档构建话题模型,最后在整个时间序列上建立话题演化关系。由于该模型对固定的时间窗口内的文档构建话题模型,因此需要考虑窗口内话题相关性的问题。此外,由于话题数目固定,使得该模型无法检测文档中的新话题。时间变量结合型直接将时间信息作为变量或特征属性引入到话题模型中,通过构建话题模型来直观反映话题随时间的演化。该模型由于考虑了文档的有序性,因此较后离散时间型更能体现话题相关文档的前后因果关系。同时,由于该模型没有马尔科夫的假设,因此不需要考虑前一个状态的模型对于当前模型的影响,较先离散时间型具有较

到稿日期:2015-05-13 返修日期:2015-06-23 本文受国家自然科学基金项目(41471371),江苏省高校自然科学基金项目(13KJB520014)资助。
徐 伟(1990—),男,硕士生,主要研究方向为数据挖掘技术与应用,E-mail:xwnjnu@163.com;赵 斌(1978—),男,博士,讲师,主要研究方向为Web数据挖掘,E-mail:zhaobin@outlook.com;吉根林(1964—),男,教授,博士生导师,主要研究方向为数据挖掘技术及应用,E-mail:glji@njnu.edu.cn(通信作者)。

少的参数推导和计算。该类关联方法认为不同话题间存在明显的时间分割点，适用于文本时间跨度大且话题内容差异明显的情况。第二种关联方法是结合文本自身特性，构建关联函数来判定不同时间段的话题是否具有相同的语义。如文献[9]利用话题的内容信息，采用 KL 距离(Kullback Leibler Divergence)来衡量话题间的关联性。该方法充分考虑了文本的内容特征，适用于文本长度长、话题信息完整的长文本。文献[10]从网页文档自身的引用特性出发，利用不同话题成员文档集之间的交叉引用次数构造关联函数，来判断两个话题的关联程度。该方法利用网页文档的非文本特征来弥补其文本特征不足的劣势，从而提高话题关联的准确度。可以发现，此类关联方法主要通过结合文本特征和非文本特征，构建关联函数来判别话题间的演化关系。

上述的话题演化主要以长文本为研究对象，而本文以微博短文本为对象，研究基于转发关系的热点事件话题演化方法。由于微博的数据特征和传播特性与传统媒体(新闻、网页文档)差异明显，因此现有的话题演化方法无法直接应用于微博，具体理由为：第一，长文本的话题演化方法不适合短文本。现有的话题演化方法主要针对新闻、网页这样的长文本，仅采用文本特征度量话题的关联性。而微博作为短文本的代表，文本长度短，且文本特征不足，仅依靠文本特征很难准确度量话题间的关联关系。第二，微博话题的演化特征与新闻不同。新闻文本按照时间顺序发布，形成序列结构。而微博是通过媒体传播的方式发布消息，按照树形结构组织。第三，两种文本的链接关系也不相同。微博的转发和网页的引用都是一种链接关系。网页文档中的引用关系存在内容信息的包含关系，一篇文档可以引用多篇不同文档。但是，具有转发关系的微博在内容上不存在包含关系，而且一条微博只能转发一条微博信息。

不难发现，微博的话题演化无法借鉴长文本的方法，应该从微博消息自身特点出发进行研究。根据此思路，本文提出了基于转发关系的微博话题演化算法(Microblog Topic Evolution based on Retweeting Relationship，MTERR)，该算法分为两个阶段：话题抽取和话题关联。在话题抽取阶段以 LDA 模型为基础，通过引入时间信息并结合微博的转发特性，遍历时间轴上的微博消息序列提取话题信息；在话题关联阶段综合考虑话题的内容信息和话题下微博消息集间的转发关系，构建出话题间的关联函数，重现微博话题的演化过程。

2 问题描述

在微博话题演化问题研究中，设微博转发图 $Gw=(Vw, Ew)$，其中 Vw 为微博消息集合，$Vw=\{w_1,\cdots,w_n\}$，其中 w_i 表示第 i 条微博，$w_i=(wc_i, wt_i, wf_i)$，wc_i 代表微博 w_i 的文本内容；wt_i 代表微博 w_i 的时间，时间的基本单位为“日”；wf_i 表示微博 w_i 的微博消息转发集，记为 $wf_i=\{w_{i1}\cdots w_{im}\}$，$Ew$ 为微博间的转发关系集合，$Ew=\{\langle w_i, w_j\rangle \mid w_i\in Vw, w_j\in Vw\}$，$\langle w_i, w_j\rangle$ 为微博 w_i 到 w_j 的转发。微博话题演化图 $Gtp=(Vtp, Etp)$，其中 Vtp 为话题集合，$Vtp=\{tp_1\cdots tp_q\}$，其中 tp_i 表示第 i 个话题，$tp_i=(tpt_i, tpw_i, tpf_i)$，$tpt_i$ 代表话题 tp_i 的产生时间；tpw_i 代表产生话题 tp_i 的起始微博；tpf_i 代表话题 tp_i 下的微博消息集。Etp 为话题间的演化关系集合。文中所有序列和集合的规模都记为 $|\cdot|$。

微博话题演化问题描述为：给定热点事件的微博转发图 Gw，生成该事件的话题演化图 Gtp。话题演化图 Gtp 要能够反映热点事件中话题随时间的发展演化过程。

3 微博话题演化算法

微博话题演化研究的主要任务是提取热点事件中的话题信息，通过构建话题间的链接关系，反映热点事件在微博平台上的发展演化过程。针对此要求，话题演化处理框架包含 5 个阶段。

(1)数据采集：根据事件的关键字和微博的时间标签收集指定热点事件的微博。

(2)预处理：微博消息的序列化处理、文本分词、停用词过滤等。

(3)话题抽取：以话题模型为基础，并结合时间信息和微博的转发特性，提取微博话题信息。

(4)话题关联：综合考虑话题的内容信息和话题下微博消息集间的转发关系，进行话题关联。

(5)结果展示：采用可视化技术展示微博话题演化过程。

3.1 话题抽取

话题抽取阶段的主要任务是提取热点事件中的话题信息。微博作为一种新的社交媒体，消息长度短且话题信息是通过微博的转发而产生。微博短文本的特性使得单一的微博消息不足以表达完整的话题信息，如果直接采用话题模型获取话题信息，会存在数据稀疏的问题，造成最终话题的不准确性。而且，话题模型本身的文档顺序无关性的假设，使得其忽略了话题演化中重要的时间因素。除此之外，微博的转发机制使得微博通过上下文来表达语义信息，新话题的产生不仅取决于话题内容的改变，也由支撑它的微博消息数量所决定。因此，仅仅依靠话题模型不能准确提取话题信息，需要结合时间信息和微博的转发特性来提高识别话题信息的准确性。

在本文中，所有微博消息采用词项集合的形式表示。微博消息 w 中的第 i 个词项记为 $e_{w,i}$，整条微博 w 的词项集记为 $E(w)$，整个数据集词项的集合记为 $\varepsilon=\sum_{w\in Vw}E(w)$，话题模型得到的话题分布集合记为 TD。

3.1.1 利用话题模型获取微博话题分布

本文选择的话题模型为 LDA 模型[5]。LDA 模型是在 PLSI 模型的基础上，使用一个服从狄利克雷(Dirichlet)分布的 K 维隐含随机变量来表示文档的话题概率分布。本文利用斯坦福大学的话题建模工具[11]对微博消息集合构建 LDA 模型。在 LDA 模型中，由于话题在词项上的分布都是多项式分布，因此对 LDA 模型求解之后可以得到话题在词项上的分布矩阵。该矩阵定义如下所示：

$$\begin{array}{c} \\ td_1 \\ \vdots \\ td_k \end{array}\begin{array}{c} \begin{array}{ccc} e_1 & \cdots & e_m \end{array} \\ \begin{bmatrix} \varphi_{11} & \cdots & \varphi_{1m} \\ \vdots & \ddots & \vdots \\ \varphi_{k1} & \cdots & \varphi_{km} \end{bmatrix} \end{array}$$

其中，φ_{ij} 表示在词项 e_j 上话题分布 td_i 的概率值，k 为 LDA 模型初始设定的话题分布数，$e_j\in\varepsilon$，$td_k\in TD$。

由于微博在转发图中起到承上启下的作用，因此可能隶属于多个话题区域，只有最佳的话题分布才最能代表该微博消息。为了找出能表示微博 w_i 的最佳话题分布，本文在 k 个话题分布中选取具有最高概率值的话题分布作为该微博 w 的最佳话题分布 $Btd(w)$，微博的最佳话题分布计算公式如下所示：

$$Btd(w)=\underset{td\in TD}{\operatorname{argmax}}\ p(w|td)$$

$$p(w|td)=\prod_{i=1}^{|E(w)|}p(e_{w,i}|td)$$

其中，$Btd(w)\in TD$，$p(w|td)$表示微博 w 在话题分布 td 下的概率，$p(e_{w,i}|td)$表示词项 $e_{w,i}$在 td 下的概率。

3.1.2 结合时间因素和转发特性抽取话题信息

在话题模型中，文档顺序无关性的假设导致文档间的顺序可以任意调换，从而忽略了文档中的时间属性。而话题作为一种数据流形式的文本抽象，随着时间推移而不断演化。热点事件发展的时间顺序决定了话题的演化顺序，因此时间因素在话题演化中不可或缺。除此之外，微博作为社交媒体的代表，与传统媒体最大的差别在于“社交”。在热点事件的微博讨论中，新话题由大量的用户观点所产生，而用户参与讨论、发表自身观点主要是通过微博的转发。一条微博的转发量越多，就越能引领一个新的话题，因此充分利用转发特性确定新话题的产生显得非常重要。

新话题产生的基本思想：首先考虑时间因素的影响，对微博消息集合 Vw 按时间先后顺序进行排序，得到微博消息序列。然后在时间轴上依次遍历微博消息序列，如果微博 w_i 的内容信息足够新颖且其转发消息集 wf_i 中的微博消息多数与其内容相近，即支持微博 w_i 的微博消息足够多，则认为微博 w_i 可以引领一个新话题。

如图 1 所示，假设前 $h-1$ 条微博生成 b 个话题 $tp_1\cdots tp_b$，第 h 条微博 w_h 的内容新颖性定义如下：

$$Nov(w_h)=\log\frac{p(w_h|Btd(tpw_1),\cdots,Btd(tpw_b),Btd(w_h))}{p(w_h|Btd(tpw_1),\cdots,Btd(tpw_b))}$$

$$p(w|td_1\cdots td_k)=\prod_{i=1}^{|E(w)|}\left(\sum_{j=1}^{k}\frac{1}{k}\cdot p(e_{w,i}|td_j)\right)$$

其中，$Nov(w_h)$表示微博 w_h 的内容新颖性，微博 w_h 的内容与前面 b 个话题越不相同，则分子与分母的差距越大，说明微博 w_h 越有可能产生新的话题，否则微博 w_h 应该隶属于前面某个话题。$p(w|td_1\cdots td_k)$表示微博 w 在话题分布 $td_1\cdots td_k$ 下的概率。

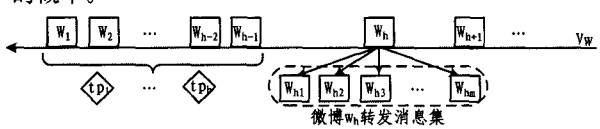


图 1 微博话题抽取示意图

新话题的产生除了内容要具有新颖性之外，还需有一定数量微博消息的支持。微博 w_h 的支持度定义如下：

$$Sup(w_h)=\sum_{w\in wf_h}\log\frac{p(w|Btd(tpw_1),\cdots,Btd(tpw_b),Btd(w_h))}{p(w|Btd(tpw_1),\cdots,Btd(tpw_b))}$$

其中，$Sup(w_h)$表示微博 w_h 的支持度，如果微博 w_h 转发消息集中的微博 w 与微博 w_h 内容越相近，且与前面 b 个话题内容相差较大，说明微博 w 支持微博 w_h。如果转发消息集中支持微博 w_h 的微博数较多，则微博 w_h 越有可能引领新的话题。

分别对微博内容新颖性和支持度设定阈值 C_N 和 C_S，只有当微博 w_h 具有新颖的内容且具有一定的支持度时，才能够形成新的话题，即满足：

$$Nov(w_h)\geqslant C_N,Sup(w_h)\geqslant C_S$$

此时生成新的话题 tp_{b+1}，$tpt_{b+1}=wt_h$，$tpw_{b+1}=w_h$，$tpf_{b+1}=wf_h\cup\{w_h\}$。

3.1.3 根据微博话题分布确定所属话题

话题抽取过程中，由于某些微博消息没有产生新的话题，使得这些微博不属于任何话题，因此需要确定这些微博消息的话题归属。由于每条微博都具有 k 个话题分布，每个话题都有代表的起始微博且起始微博的最佳话题分布也属于 k 个话题分布中，因此只需要将微博的 k 个话题分布按分布概率进行排序，然后与话题起始微博的话题分布进行比较，找出第一个存在于话题中的话题分布，则该微博就隶属于该话题分布的话题。

假设 $k=5$，话题模型得到的话题分布集合 $TD=\{td_1,td_2,td_3,td_4,td_5\}$，经过话题抽取得到两个话题$\{tp_1,tp_2\}$，且两个话题起始微博的最佳话题分布分别为 $Btd(tpw_1)=td_1$，$Btd(tpw_2)=td_3$。首先对微博 w_i 的 5 个话题分布按分布概率由高到低进行排序，得到的话题分布为$\{td_2,td_3,td_5,td_1,td_4\}$。然后遍历微博 w_i 排序后的话题分布，可以发现第一个存在于话题中的话题分布为 td_3，所以微博 w_i 属于话题 tp_2，并将微博 w_i 加入话题 tp_2 的微博消息集中，$tpf_2=tpf_2\cup\{w_i\}$。

3.2 话题关联

话题关联阶段的主要任务是对话题抽取阶段所得到的话题信息进行关联，重现话题的演化过程。由于微博消息文本长度短、信息量不足，只考虑话题的文本特征可能导致以下两个问题：1)具有相同语义但内容相差较大的微博不能被正确关联；2)具有相近内容但语义不同的微博被错误关联。因此，仅利用话题的文本特征不足以准确判定两个话题之间的关联性，还需要考虑非文本特征。

本文综合考虑话题的内容信息和话题下微博消息集间的转发关系，提出话题内容相关性和话题转发关联度两个指标，以构建话题关联函数，实现话题关联。

3.2.1 话题内容相关性指标

在热点事件的发展中，如果两个话题存在演化关系，则在内容上应该具有衔接关系，两者的语义信息应该较为接近。因此，判断话题是否关联需要考虑其内容信息。内容相关性指标是指两个话题在内容上的相关程度，是衡量两个话题是否具有演化关系的一个重要指标。内容相关程度越大表明两个话题所表达的语义信息更为接近，则该两个话题就越可能存在演化关系。由于话题的起始微博是话题的代表，最能表示话题的内容信息，因此利用起始微博间的内容相似性就可以度量话题间的内容相关性。

本文的内容相关性计算基于文献[9]中的 KL 距离，它是用来衡量两个话题间的相似程度，其公式为：

$$KL(tp_i,tp_j)=\sum_{e\in\varepsilon}p(e|Btd(tpw_i))\cdot\log\frac{p(e|Btd(tpw_i))}{p(e|Btd(tpw_j))}$$

$$KL(tp_j,tp_i)=\sum_{e\in\varepsilon}p(e|Btd(tpw_j))\cdot\log\frac{p(e|Btd(tpw_j))}{p(e|Btd(tpw_i))}$$

其中，$KL(tp_i,tp_j)$表示话题 tp_i 到 tp_j 的 KL 距离，$KL(tp_j,tp_i)$表示话题 tp_j 到 tp_i 的 KL 距离。

话题 tp_i 和 tp_j 的内容相关性 $CR(tp_i,tp_j)$定义为：

$$CR(tp_i,tp_j)=\frac{1}{2}\cdot(KL(tp_i,tp_j)+KL(tp_j,tp_i))$$

其中，$tpt_i<tpt_j$。

3.2.2 话题转发关联度指标

微博与传统媒体最大的差别在于“社交”。在热点事件的

微博讨论中，大量的微博消息由用户间的互动产生，而用户参与讨论的形式主要是通过微博的转发。伴随事件持续发展，大量个人意见和评论在微博平台上逐渐汇聚融合形成群体话题。由此可见，在微博平台上，热点事件中的话题是通过微博的转发而传播，话题间的演化关系应该蕴含在微博的转发中。因此，话题间的演化需要考虑微博间的转发关系。

话题转发关联度指标是指两个话题下微博消息集间的转发情况。如果话题 tp_j 从话题 tp_i 演化而来，则属于话题 tp_j 的大量微博消息应该从属于话题 tp_i 的微博消息转发而来。因此利用两个话题下微博消息集间的转发关系可以判定两个话题是否关联。

话题 tp_i 和 tp_j 的转发关联度 $FR(tp_i,tp_j)$ 定义为：

$$FR(tp_i,tp_j)=\sum_{w_a\in tpf_i}\sum_{w_b\in tpf_j}|\{\langle w_a,w_b\rangle|\langle w_a,w_b\rangle\in Ew\}|$$

其中，$tpt_i<tpt_j$。

3.2.3 *构建话题关联函数*

为了从上述两个方面综合考虑话题间的关联性，本文采用线性组合的方法定义了两个话题间的关联函数 $F(tp_i,tp_j)$，其公式为：

$$F(tp_i,tp_j)=\lambda\cdot\frac{CR(tp_i,tp_j)}{\max\limits_{tp_a\in Vtp,tp_b\in Vtp}CR(tp_a,tp_b)}+(1-\lambda)\cdot\frac{FR(tp_i,tp_j)}{\max\limits_{tp_a\in Vtp,tp_b\in Vtp}FR(tp_a,tp_b)}$$

本文设定关联度阈值 ω，如果话题 tp_i 和 tp_j 的关联函数值 $F(tp_i,tp_j)>\omega$，则话题 tp_i 和 tp_j 相互关联，话题 tp_i 到 tp_j 具有演化关系，否则不是。由于无法判定话题内容相关性和转发关联度的权重大小，因此将 λ 设为 0.5。

4 实验与结果分析

4.1 话题演化评价方法

现有的评价方法主要是通过人工分析话题演化图的合理性来判别方法的好坏，并没有明确的评测指标。除此之外，由于不同方法生成的演化图中话题内容和关联边的条数都各自不同，因此很难从整体上评价演化方法的优劣。为了评价微博话题演化的正确性，本文采用人工评测的方法，分别从演化图的点、边和子图 3 个方面进行评测。

在点的评价方面，主要采用信息检索中的准确率、召回率和 F1 值 3 种指标[12]评价微博话题演化中的话题正确情况和全面程度，并对每个话题进行正确性判定，从而得到各个算法在话题抽取阶段的正确率；在边的评价方面，主要针对演化图中的每一条边的合理性进行分析，通过判断每一条边的正确性，得到各个算法在话题关联阶段的正确率；在子图的评价方面，主要是对完整演化图进行分割，形成多个子图，然后依据事件的发展情况，对子图中话题演化的合理性进行分析。

4.2 实验结果与分析

为了研究微博热点事件的话题演化问题，本文采用腾讯微博 API 收集了“马航 MH370 失联”事件（MH370）的微博消息，选取最大的转发连通图作为实验数据集。该数据集的时间跨度为 2014 年 3 月 15 日到 2014 年 4 月 18 日，微博总条数为 3129 条，数据集大小为 1172kB。

实验中采用人工抽取的方式从 MH370 实验数据集中识别话题信息（即人工摘要）。为了尽可能降低人的主观因素对人工摘要结果的影响，本文将 MH370 事件的新闻时间线[13]作为人工摘要的指导信息，然后由 5 名不同研究人员分别对 MH370 实验数据集进行人工摘要，得到 5 份结果。如果在特定时间段内某条微博消息出现在 4 份或 5 份的摘要结果中，则此微博应该被选入最终的人工摘要中。按照该方法，最终的人工摘要包含 17 条微博。

为了验证微博话题演化算法的可行性与有效性，本文选取文献[10]中的话题拓扑结构（Topology of Topic Evolution，TTE）算法作为基准测试算法。参与微博话题演化实验的两种算法参数设置如下：TTE 算法中混合模型的话题分布数设为 17，一元混合模型和背景模型的权重均设为 0.5，关联函数阈值设为 2。MTERR 算法中 LDA 模型的超参数 $\alpha=50/k$，$\beta=0.01$，话题分布数 $k=17$，内容新颖性 C_N 和支持度阈值 C_S 分别设为 5 和 4，关联函数阈值 ω 设为 0.55。为了保证两种算法比较的公平性，本文使得两种算法最终均生成 17 条话题信息。该实验在 CPU 为 Intel Core i5 3.2GHz、内存大小为 4GB 的 PC 机上进行。

在准确率、召回率和 F1 值指标方面采用词项粒度对两种算法进行了性能比较，实验结果如表 1 所列。表 2 为两种算法在话题抽取阶段的正确率。表 3 为两种算法在话题关联阶段的正确率。

表 1 MH370 实验数据集上两种算法的准确率、召回率和 F1 值比较

算法	词项粒度		
	准确率	召回率	F1 值
TTE	67%	58%	62%
MTERR	78%	72%	75%

表 2 MH370 实验数据集上两种算法在话题抽取阶段的正确率

算法	话题总数	正确个数	正确率
TTE	17	12	71%
MTERR	17	15	88%

表 3 MH370 实验数据集上两种算法在话题关联阶段的正确率

算法	边总数	正确个数	正确率
TTE	31	17	55%
MTERR	33	23	70%

通过设定关联函数阈值对边进行过滤，从而将整个演化图分成多个子图。为了确保两个算法比较的公平性，对阈值进行调整，使得两个算法得到的演化子图集合中均保留 17 条边。

在 TTE 算法中，完整演化图分成 5 个子图，如图 2 所示，总共存在 29 条演化路径，分别对该 29 条演化路径的合理性进行分析，使用“合理”、“部分合理”和“不合理”3 种情况对其进行评判，“合理”表示该演化路径完全可解释，“部分合理”表示该演化路径中只存在部分话题演化可解释，“不合理”表示该演化路径完全不可解释，结果如表 4 所列。

表 4 TTE 和 MTERR 算法话题演化子图的合理性分析结果

算法	演化路径总数	最大演化深度	合理数	部分合理数	不合理数	合理率
TTE	29	4	9	18	2	31%
MTERR	27	9	15	12	0	56%

在 MTERR 算法中，完整演化图分成 3 个子图，如图 3 所示，总共 27 条演化路径，分别对该 27 条演化路径的合理性进行分析，同样使用“合理”、“部分合理”和“不合理”3 种情况对其进行评判，结果如表 4 所列。

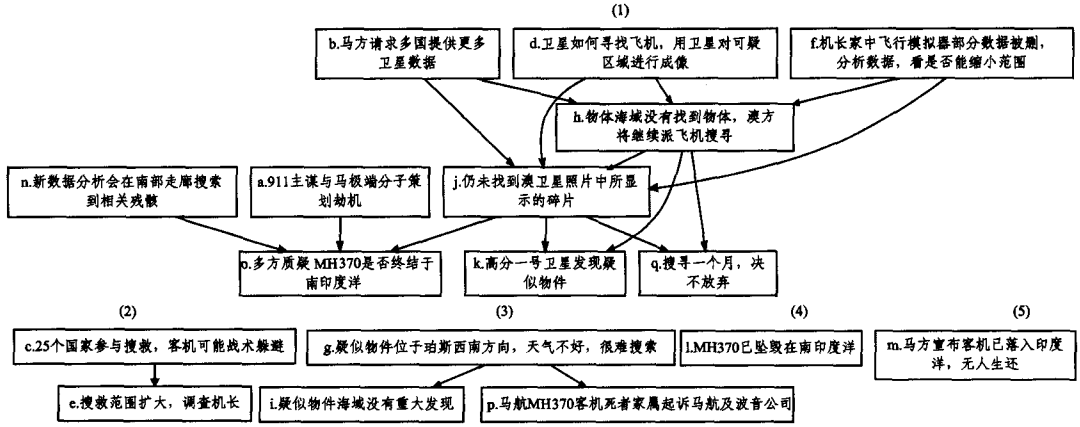


图 2　TTE 算法的话题演化子图

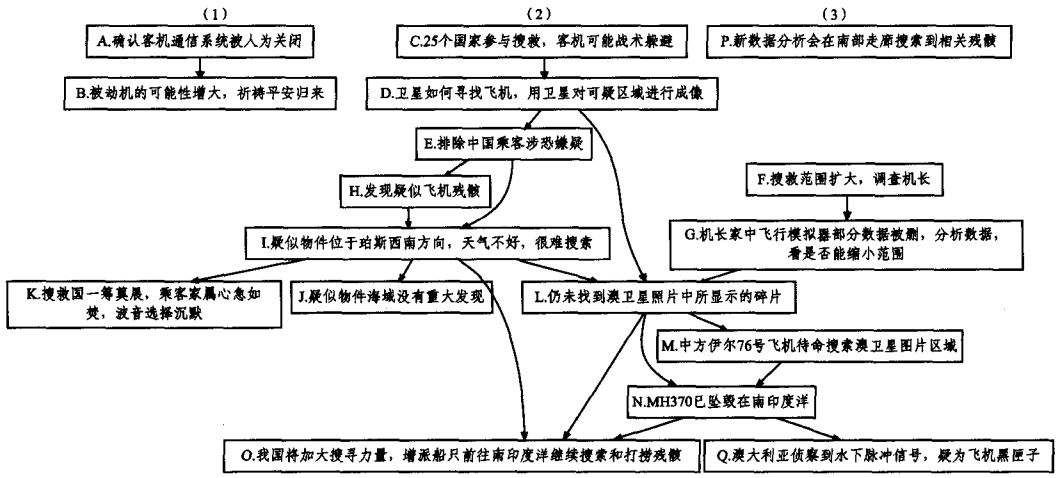


图 3　MTERR 算法的话题演化子图

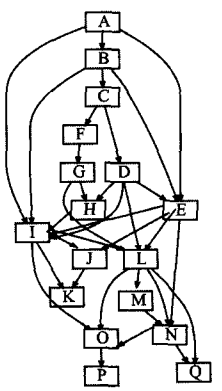
从上述实验结果可以发现：

在话题抽取的准确性方面，MTERR 算法优于 TTE 算法。主要原因是 MTERR 算法以 LDA 模型为基础，综合考虑时间因素和微博转发特性提取话题信息。而 TTE 算法采用混合模型进行话题抽取，仅利用文档的文本特性，而微博消息长度短且文本特征不足，使得该方法实验效果不理想。

在话题关联的正确性方面，MTERR 算法效果较好。原因是，TTE 算法在话题关联中只利用话题成员文档集之间的相互引用关系来判定两个话题是否关联。这样会出现两个语义相同的话题由于成员文档集之间没有引用关系而不能被正确关联的情况。而 MTERR 算法不仅利用了文档间的引用关系，而且也考虑了话题间的内容相关性，因此关联的结果比 TTE 算法好。

在演化子图的合理性方面，MTERR 算法优于 TTE 算法。首先 MTERR 算法中子图的最大演化深度比 TTE 算法深，演化深度越深，说明话题间的关联性越好，话题演化的时间跨度越长，话题的发展趋势越明显。其次 MTERR 算法虽然得到的演化路径总数比 TTE 算法少，但是合理数却比 TTE 算法多，而且不存在完全不合理的演化路径，表明 MTERR 算法得到的话题演化图具有更好的可解释性。

为了展示 MTERR 算法生成的话题演化图的具体效果，图 4 展示了该算法生成的 MH370 事件的话题演化图。由于篇幅原因，本文只用简短的语句来表达话题信息。



编号	话题信息	编号	话题信息
A	确认客机通信系统被人为关闭	J	疑似物件海域没有重大发现
B	被劫机的可能性增大，祈祷平安归来	K	搜救国一筹莫展，乘客家属心急如焚
C	25 个国家参与搜救，客机可能战术躲避	L	仍未找到澳卫星照片中所显示的碎片
D	如何寻找飞机，用卫星对可疑区域进行成像	M	中方伊尔 76 号飞机待命搜索澳卫星图片区域
E	排除中国乘客涉恐嫌疑	N	MH370 已坠毁在南印度洋
F	搜救范围扩大，调查机长	O	我国将加大搜寻力量，增派船只前往南印度洋继续搜索和打捞残骸
G	机长家中飞行模拟器部分数据被删，分析数据，看是否能缩小范围	P	新数据分析会在南部走廊搜索到相关残骸
H	发现疑似飞机残骸	Q	澳大利亚侦察到脉冲信号，疑为飞机黑匣子
I	疑似物件位于珀斯西南方向，天气不好很难搜索		

图 4　MTERR 算法的话题演化图

（下转第 100 页）

易，并且能大大缩短正则化参数以及高斯核参数的优化时间。同伦正则化思想还可以应用到深度学习等其他的研究领域中。

参 考 文 献

[1] Vapnik V. Statictical learning Theory[M]. New York: Wiley, 1998

[2] Zhang Rui, Ma Yi-chen. Kernel Methods for Pattern Analysis [D]. Xi'an: Xi'an Jiaotong University, 2009(in Chinese)
张瑞，马逸尘. 模式分析的核方法[D]. 西安：西安交通大学，2009

[3] Zhang R, Wang W J, Ma Y C. Least square transduction supportvector machine[J]. Neural Processing Letters, 2009, 29(2): 133-142

[4] Ye N, Sun R, Liu Y, et al. Support vector machine with orthogonal Chebyshev kernel[C]// Proceedings of the 18th international conference on Pattern Recognition. 2006

[5] Sedat O, Chen C H, Hakan A C. A set of new Chebyshev kernel function for support vector machine pattern classification[J]. Pattern Recognition, 2011, 44(7): 1435-1447

[6] Zhang R, Wang W J. Facilitating the Applications of Support machine by Using a new kernel[J]. Expert Systems with Application, 2011, 38: 14225-14230

[7] Zhang Rui, Gao Hong, Zhang Li-wei. A New Set of Hemite Kernel Functions for Support Vector Machine[J]. Journal of Shanxi University(Natural Science Edition), 2012, 35(1): 38-42(in Chinese)
张瑞，高红，张立伟. 一类新的支持向量机核函数-埃尔米特核函数[J]. 山西大学学报(自然科学版)，2012，35(1)：38-42

[8] Zhang Rui, Wang Wen-jian, Zhang Ya-dan, et al. Legendre Kernel Function for Support Vector Classification[J]. Computer Science, 2012, 39(7): 222-224(in Chinese)
张瑞，王文剑，张亚丹，等. 基于支持向量机分类问题的勒让德核函数[J]. 计算机科学，2012，39(7)：222-224

[9] Zhang Rui, Wang Wen-jian, Wang Jia-qi, et al. Laguerre Kernel Functions for Support Vector Classification[J]. Computer Engineering and Applications, 2012, 48(36): 50-53(in Chinese)
张瑞，王文剑，王嘉琦，等. 一类新的基于拉盖尔正交多项式的核函数[J]. 计算机工程与应用，2012，48(36)：50-53

[10] Zhang Rui, Yang Xiao, Tan Xiu-lin. New SVM Kernel Function Based on Gegenbauer Polynomial[J]. Journal of Shanxi University(Natural Science Edition), 2013, 36(1): 30-33(in Chinese)
张瑞，杨晓，谭秀林. 基于盖根鲍尔多项式的 SVM 核函数[J]. 山西大学学报(自然科学版)，2013，36(1)：30-33

[11] Liao Shi-jun. Beyond the Perturbation: the basic idea and application of the Homotopy analysis method [J]. Journal of Mechanics, 2008, 38(1): 1-34

[12] Agarwal R P, O'Regan D. Homotopy and existence of solutions for the nonlinear equation $Lx \in Nx$[J]. Nonlinear Analysis, 2001, 44(4): 537-544

[13] Soriano J M. On the existence of zero points of a continuous founction[J]. Acta Mathematica Scientia, 2002, 22(2): 171-177

(上接第 82 页)

结束语 本文以微博消息为对象，研究微博热点事件中的话题演化问题。按照处理流程的不同，微博话题演化处理框架分为 5 个阶段，分别是数据采集、预处理、话题抽取、话题关联和结果展示。在话题抽取阶段，以 LDA 模型为基础，结合时间因素和微博转发特性，提取微博话题信息。在话题关联阶段，综合考虑话题的内容信息和话题下微博消息集间的转发关系，通过话题内容相关性和话题转发关联度两个指标来构建话题关联函数，重现话题演化过程。利用真实微博数据进行实验，结果表明 MTERR 算法可以有效地抽取话题信息，构建话题演化图，从而反映热点事件发展演化的过程。

参 考 文 献

[1] Hong Yu, Zhang Yu, Liu Ting, et al. Topic Detection and Tracking Review[J]. Journal of Chinese Information Processing, 2007, 21(6): 71-87(in Chinese)
洪宇，张宇，刘挺，等. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报，2007，21(6)：71-87

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

[3] Kumaran G, Allan J. Text classification and named entities for new event detection [C]// Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield: ACM, 2004: 297-304

[4] Hofmann T. Probabilistic Latent Semantic Indexing [C]// Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley, CA, USA: ACM, 1999: 50-57

[5] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022

[6] Ha-Thuc V, Mejova Y, Harris C G, et al. A relevance-based topic model for news event tracking [C]// Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Boston, MA, USA: ACM, 2009: 764-765

[7] Alsumait L, Barbar D, Domeniconi C. On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking [C]// Proceedings of the 8th IEEE International Conference on Data Mining. Pisa, Italy: IEEE Computer Society, 2008: 3-12

[8] Wang X, Zhang K, Jin X M, et al. Mining common topics from multiple asynchronous text streams [C]// Proceedings of the Second International Conference on Web Search and Web Data Mining. Barcelona, Spain: ACM, 2009: 192-201

[9] Gohr A, Hinneburg A, Schult R, et al. Topic Evolution in a Stream of Documents [C]// Proceedings of the SIAM International Conference on Data Mining. Sparks, Nevada, USA: SIAM, 2009: 859-870

[10] Jo Y, Hopcroft J E, Lagoze C. The web of topics: discovering the topology of topic evolution in a corpus [C]// Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India: ACM, 2011: 257-266

[11] http://nlp.stanford.edu/software/tmt/tmt-0.4

[12] Manning C D, Raghavan P, Schutze H. Introduction to Information Retrieval [M]. New York: Cambridge University Press, 2008: 142-143

[13] MH370 [EB/OL]. (2014-03-08)[2014-10-30]. http://baike.baidu.com/view/12368712.htm