论坛主题挖掘研究综述

陈 迪,代艳君,王志锋
CHEN Di, DAI Yanjun, WANG Zhifeng

华中师范大学 教育信息技术学院,武汉 430000
School of Educational Information Technology, Central China Normal University, Wuhan 430000, China

CHEN Di, DAI Yanjun, WANG Zhifeng. Survey of research on forum topic mining. Computer Engineering and Applications, 2017, 53(16):36-44.

Abstract: With the advent of the big data age, network forum data which is social, randomness and decentralized is exploding and difficult to be used directly. Forum topic mining can refine the main forum argument yet. It can identify the content of the user's discussion from the complex forum data and extract the theme. This paper describes the problem and the framework of the forum topic mining, and classifies of existing technologies, basic types as forum text preprocessing, topic mining algorithm and topic modeling. Then, the basic characteristics and typical methods of the above three kinds of topic mining technology are described, compared and summarized in detail. At the end of the paper, discusses and analyzes the current problems and development trend of the forum topic mining.

Key words: forum mining; topic mining; text preprocessing; topic model

摘 要: 伴随着互联网大数据时代的来临,网络论坛数据呈爆炸式增长,这类数据具有社会性、随意性、分散性等特点,难以被直接使用。而论坛主题挖掘技术能从复杂的论坛数据中识别出用户集中讨论的文本内容,并从中提取主题,以达到提炼论坛主要论点的目的。对论坛主题挖掘进行了问题描述和任务框架梳理,并依照任务框架对现有技术进行了分类,基本类型为论坛文本预处理、主题挖掘算法和主题建模,详细阐述了以上三类论坛主题挖掘技术的基本特征和典型方法,进行了比较与总结,对论坛主题挖掘当前存在的问题及其发展趋势进行了分析与讨论。

关键词: 论坛挖掘;主题挖掘;文本预处理;主题模型
文献标志码: A **中图分类号:** TP391 doi:10.3778/j.issn.1002-8331.1705-0183

1 引言

随着信息技术的飞速发展,供人们交流沟通的虚拟空间应运而生,论坛作为一种依托于互联网的典型虚拟互动社区,已经成为日常生活中不可或缺的一部分。论坛允许用户自主开贴、自由回复,所产生的讨论内容信息量巨大,既包含了用户的广泛观点,也反映了用户的关注焦点,但论坛用户发言的随意性会导致大量噪声数据的产生,如错误表达或无意义内容,另外,论坛帖子依据时间先后顺序排列,内容接近的文本可能在网页位置上相距甚远,因此,论坛的主要论点无法直接获得,且随着论坛数据量的激增,论点数量随之增长,由此论坛主题挖掘技术应运而生。论坛主题挖掘技术从论坛数据中识别出具有主题相关性的内容,并从中提取主题,该技术能获取论坛网站中的主题分布情况或沿时间线的主题演化情况。

事实证明,论坛主题挖掘具有重要意义。如对热点话题的识别或对突发话题的检测可有效应用于网络舆情检测[1-2],而高质量话题的抽取或指定话题的抽取可有效应用于论坛信息检索[3]、用户行为分析[4-5]等领域。

2 论坛数据挖掘研究框架

信息时代来临,现代网民越来越热衷于在网络论坛中交流互动,同时,论坛主题挖掘也成为了一个受到广泛关注的研究方向。

2.1 问题描述

一般认为论坛主帖是发起话题的第一个帖子,在主帖下回复的帖子称为跟帖,通常将主帖及其所有跟帖的组合称为线程,而网络论坛则是由一系列线程构成的。

基金项目: 国家自然科学基金(No.61501199);国家科技支撑计划(No.2015BAK33B02)。
作者简介: 王志锋(1985—),男,博士,讲师,研究领域为信号处理、机器学习与数据挖掘,E-mail:zfwang@mail.ccnu.edu.cn。
收稿日期: 2017-05-15 **修回日期:** 2017-06-27 **文章编号:** 1002-8331(2017)16-0036-09

由于论坛发言的随意性，跟帖常偏移于主帖内容而任意展开，主帖内容有时并非线程的主题。

主题挖掘任务的本质是将输入的文本流划分到不同的主题类中，并且在必要时候建立新的主题类。由于论坛特殊的交流模式，其文本数据具有以下特点：(1)口语化，帖子发布者来自不同的地方，也有着不同的经历与背景，在表达同一观点时措辞会有很大不同；(2)篇幅差距较大，有些帖子的论述较多，而有些帖子只是一些短语或词语；(3)存在许多不规范甚至错误的表达方式；(4)论坛中有大量未在字典中列出的新词，且这类词的数量正在日益增长。这些特性为论坛主题挖掘工作带来一定的挑战。

2.2　任务框架

论坛主题挖掘任务除基本的论坛话题识别外，还包括热点话题检测、突发话题检测、高质量话题抽取、指定话题抽取等等。其基本任务框架如图1所示，论坛主题挖掘的数据处理对象为帖子文本，首先需通过网站爬虫技术或在开源论坛数据库中获取论坛文本流，然后对论坛文本进行预处理，过滤其中的无效数据，接下来通过主题挖掘算法提取用户集中讨论的内容，再通过主题建模完成对以上内容的主题描述，最终生成主题。

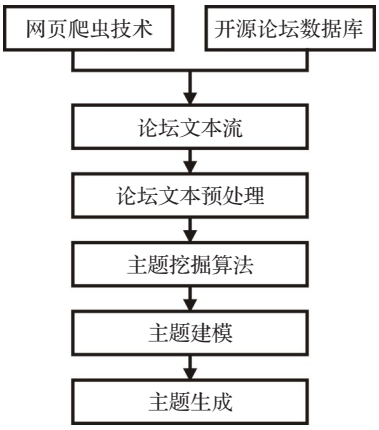

图1　论坛主题挖掘任务框架

3　论坛主题挖掘技术

论坛主题挖掘技术的目标是从论坛文本流中检测出用户集中讨论的内容，且用较短文本描述它们，从而生成讨论主题。根据论坛主题挖掘任务框架，可将论坛主题挖掘技术分为论坛文本预处理、主题挖掘算法和主题建模三种类型。

3.1　论坛文本预处理

作为论坛主题挖掘的第一步，论坛文本预处理的目的是过滤原始论坛数据中的无效数据，同时将文本转换为便于计算机处理计算的数据对象。

论坛文本预处理过程通常遵循以下步骤，如图2所示：① 对文本进行分词；② 去除其中的停用词[6]，停用词是指没有实际含义的字词，每种语言都有对应的停用词表，且较为固定；③ 进行词频统计，即每个词出现的频率；④ 进行文本向量化，将文本数据转换为易于数学处理的向量形式。

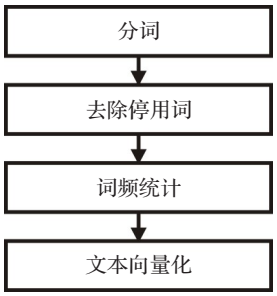

图2　论坛文本预处理的一般步骤

上述步骤中，常用的分词工具包括ICTCLAS[7]、Ansj、SCWS、FudanNLP[8]、Stanford[9]、mmseg4j、CRF++[10]等，基本信息及原理如表1所示。上述分词工具适用于所有类型文本，但对论坛数据而言，由于其用户具有年轻化趋向，不规范的表达或新词层出不穷，且不排除这些新词正是讨论热点的可能性，因此对于论坛主题挖掘，分词工具存在一定的缺陷。所以在某些论坛主题挖掘研究工作中，为避免新词遗漏，会明确规定需手动添加新词，但其完整性仍无法保证。Li等[11]为解决这一问题，提出了一套基于最长公共分段连续子序列LCSCS(Longest Common Segmented Consecutive Subsequence)算法。该算法依据热门主题所具有的三个特性：① 被大量讨论；② 可从帖子标题提取；③ 可从帖子标题中提取字符序列。通过提取有效的标题内容，并检测出分段连续子序列，完成文本分词。虽然该方案仅针对论坛标题进行处理，对于热门主题检测而言有失严谨，但在论坛新词的识别上有所突破。

第④步文本向量化是将文本进行空间向量化，用数学上的多维特征向量来表示文本，以便用于后续的数

表1　常用的分词工具

分词工具	作者	语言	基本原理	准确率
ICTCLAS	中国科学院计算技术研究所	C/C++	层叠隐马尔可夫模型、原子切分、N-最短路径粗切分	98.45%
Ansj	nlpcn.org	Java	ICTCLAS的java实现	96%以上
SCWS	Hightman	C语言	词频词典的机械中文分词引擎	90%~95%
FudanNLP	复旦大学	Java	统计与规则	70.3%~98.40%
Stanford	斯坦福大学自然语言处理组	Java	条件随机场模型、字符身份特征	94.3%~96.4%
mmseg4j	chenlb	Java	正向最大匹配	98.41%
CRF++	上海交通大学	Java	条件随机场模型	96.1%~98.2%

据挖掘算法或主题模型。通常使用基于向量空间模型VSM(Vector Space Model)[12]的方式,将文本空间看作是由一组正交词条矢量所组成的矢量空间,每段文本 d 用一个范化矢量 $V(d)=(t_1,w_1(d),t_2,w_2(d),\cdots,t_n,w_n(d))$ 表示,其中 t_i 为词条项,$w_i(d)$ 表示词条 t_i 在文本 d 中的权值,用于显示向量 t_i 在文本 d 中的重要程度。可以将文本 d 中出现的所有词条作为 t_i,也可以要求 t_i 是 d 中出现的所有短语,从而提高内容特征表示的准确性。$w_i(d)$ 一般被定义为词条 t_i 在文本 d 中的出现频率 $tf_i(d)$ 的函数,$w_i=f(tf_i(d))$,这也是第③步要进行词频统计的原因,常用的 f 函数有布尔函数、平方根函数、对数函数、TF-IDF函数等。目前使用最广泛的是TF-IDF函数,其中文本内频率TF(Term Frequency)是指一个特征项在文档中出现的次数,反映了该特征项对文档的重要性,而逆文本频数IDF(Inverse Document Frequency)是指一个特征项在其他文档中的重要程度,假设某个词出现的频数越小,它区分不同类别的能力就越大,因此IDF与该词所在文档的总数成反比或近似反比。

由于论坛文本不同于文档文本,相邻文本段之间不具有明显的逻辑关系,因此预处理结果通常存在特征稀疏的问题,因此必要时可在上述论坛文本预处理步骤中加入关键词共现技术[13]或依存句法分析来应对特征稀疏的问题。

3.2 主题挖掘算法

完成论坛文本预处理后,需从得到的数据中提取出用户集中讨论的内容,即主题相关数据。为达到目的,首先要将未被集中讨论的内容清除,再将结果数据按照所描述主题的不同分离开来,当有必要提取热门主题时,还需进行热度计算。上述过程中,将会用到各种不同的主题挖掘算法。

预处理得到的数据中,不免掺杂着一定量因未受到关注而无法形成主题的内容,主题特征提取技术可将上述主题无关数据与主题相关数据分离,并能提取出具有特定特征的主题相关数据。传统的主题特征提取方法,如基于N-gram的文本块特征提取[14]、基于向量空间模型(VSM)的主题特征提取[15]和基于关联规则和粗糙集的主题特征提取[16]等方法,仅对文本进行处理。而由于论坛结构的特殊性,其标题、帖子、用户、时间等关键要素均会因主题变化而形成具有规律的特征,因此针对论坛的主题特征提取技术并非仅对文本进行处理,而是将上述要素结合并提取其特征,将具有特定主题特征的数据与其他数据分离。Chen Y等[17]利用小波变换WPT(Wavelet Packet Transform)实现高品质主题的特征提取,其中高品质主题被定义为有完整活动或事件且包含有价值内容的话题,该方案首先提取主题中的时间序列信号,并利用小波变换从时间序列信号中获取特征,再利用反向传播神经网络结合以上特征识别出高质量的讨论内容。随后,Chen Y等[2]添加了新的特征要素,提出了识别论坛突发主题的特征提取方案,首先将线程中的标题、帖子、用户等要素加权并抽象化为特征轨迹,然后将特征轨迹划分为频率段,在观察了许多特征轨迹后发现一个规律,即在爆发之前或之后,存在一段不连续的轨迹,而在突发期间,轨迹是连续的,于是使用两个参数Sumseg和Devseg来描述每段轨迹,Sumseg描述事件权重的绝对强度,Devseg描述事件权重的相对强度的平均值。将Sumseg>110且Devseg>0.3的轨迹作为突发条目,达到突发条目与非突发条目分离的目的。以上两种方法均集中关注特征要素的全局特征,与此不同的是Chen F等[18]将每个帖子定义为一个文档,用文档内容、时间戳和时间跨度等属性来描述文档的相似度,并将相似度特征分为两类:局部特征和全局特征,按照此相似度特征区分主题相关内容与主题非相关内容。文献[17]和文献[18]主要依据时间信号的特征来识别主题相关内容,而文献[2]结合了标题、帖子、用户等多种要素,且使用不同参数来衡量突发特性,以检测论坛中的突发讨论内容。此差异正体现了论坛主题特征提取技术的发展,前两者特征较单一,而后者选用了多重特征加权计算且进行了抽象化分析。

论坛通常存在多个主题,在获取主题相关数据后,需按照所描述内容的不同,将数据划分到不同的主题集合中,该过程为主题文本聚类。论坛主题挖掘的文本聚类技术最初由传统的热点发现文本聚类技术演变而来,具有代表性的是Wang等[19]在传统的文本聚类技术基础上,考虑到了论坛以用户为中心的结构特点,先对论坛结构网络进行循环分解,然后对这些帖子进行文本聚类,该方法在一定程度上提高了传统文本聚类的效率,但由于每次循环分解都将导致大量内容遗失,因此无法保证最终所得主题的代表性。为避免这一缺陷,Ma等[3]实现了一种限定主题类型的聚类方法,该方法将主题类型限定为学科知识点,将学科的特定名词和专业术语融入到基础语料库中作为分词依据,并通过分析这些关键词的频率和权重,设计了计算帖子与关键词之间相关度的算法,再以此进行主题聚类,该方法的特色在于将知识点作为关键词权重的依据,以达到帮助学习者从论坛中快速获取所需内容的目的。为沿时间线有序展示论坛主题,Zhang等[1]提出了一种基于密度的主题词聚类模型,该模型按时间跨度将帖子的主贴标题划分为多个标题集,然后对标题集进行分词,保留候选特征词并计算其权重,选取权重最大的特征项作为主题中心进行聚类,此模型仅处理主帖标题,虽因此减少了大量噪声数据,但由于论坛中常出现主题漂移的现象(即实际主题脱离主帖标题或是偏向主帖标题的某一子题),检测出的主题会不够准确。同样为跟踪论坛主题随时间的变化情况,Zhang等[20]将聚类任务分为初始步和增量步两

步执行，初始步将文本形式化为特征空间的加权特征向量，依据关键词的相似度生成主题向量集合，然后在所得的主题集合基础上将新增的数据分派到最合适的主题集合中，其优点是当有新数据出现时，无需对所有数据进行再次处理，而只需按增量步算法处理即可，而缺点是新数据可能形成的新主题将无法被检测到。经典的聚类方法包括*K*-means算法[21]、EM算法[22]、DBSCAN算法[23]、BIRCH 算法[24]、Cure算法[25]、CLARANS算法[26]等等，除文献[1]直接使用了经典的DBSCAN算法外，其他文献均创新性地提出了专门应用于论坛主题挖掘的聚类方法。文献[3]为提高聚类效率，对传统的空间向量模型VSM（Vector Space Model）进行降维，提出了一种以专业术语为特征项的文本聚类算法，文献[19]实现了一种用于循环分解论坛结构网络的层次聚类方法，文献[20]分别设计了用于主题分类的初始聚类算法和基于分类向量的增量聚类算法。总体而言，文本聚类技术能把主题相关数据划分到不同集合中，但从各个集合中提取主题是文本聚类技术所无法实现的，当数据量较大时，必须借助其他技术自动化实现。

随着论坛数据的爆炸式增长，论坛主题数目激增，挖掘出其中受到广泛关注的热门主题具有重要意义，因此诸多学者提出了一系列主题热度计算的方法。主题热度计算主要分为两种类型，一种是对帖子进行热度计算，热度值较高的帖子所属的主题即为热门主题，第二种是对每个主题集合进行热度计算，以确定热门的主题集合。较为基本的主题热度计算方法是为帖子构建热度指标体系，并对各项指标赋予权重，计算帖子的热度量化值。如梅等[27]对论坛帖子的浏览数量、回帖数量、帖子内容、会员评价、作者属性、回帖内容、回帖人属性、发表时间等共有属性进行抽取并赋予权重，来计算每个帖子的热度值，该方案的基本实现步骤如图3所示。除此之外，热度的熵值计算也是一种使用较为广泛的主题热度计算方法，Sun等[28]综合考虑话题关注度、用户参与度和发帖人因素来定义了帖子的热度熵公式，具体流程如图4所示，然后根据帖子在一段时间内的回复数量来设定熵值的阈值，选取大于阈值的帖子作为热门帖子。不同于前两者将帖子的热度值作为其所属主题热度的判断依据，谌等[29]综合考虑各主题集合的帖子篇数与帖子关注度，提出了一个基于相对熵的语义距离计算方法来评估所有的主题集合。另外，还有一些学者尝试将经典的算法或模型引入到网络社区的主题的热度计算中，如陈等[30]将每个帖子的点击数和回复数作为热度影响因子，提出了基于KNN近邻算法和夹角余弦算法来设计主题热度的评估算法，杜等[31]将点击量、评论量、文章数量和来源数量作为热度影响因子，设计了一套基于因果模型的热度计算公式。虽然上述文献都使用了主题热度计算的方式来检测热度，但是思路上略有不同。文献[27-28，30]通过热度计算来检测热度值较高的帖子，而文献[29-31]则是针对各个主题集合进行热度计算。前者适用于具有突出热门话题的论坛，而对于帖子的热度值差距不显著的论坛而言，该思路不具备较强的说服力；后者对论坛主题的热度评估在研究思路上更具整体性和可靠性。

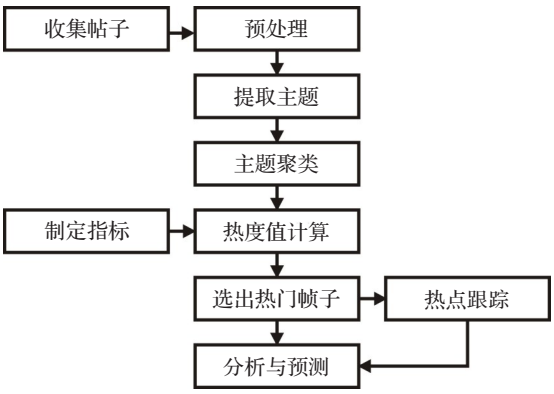

图3　热帖检测流程图

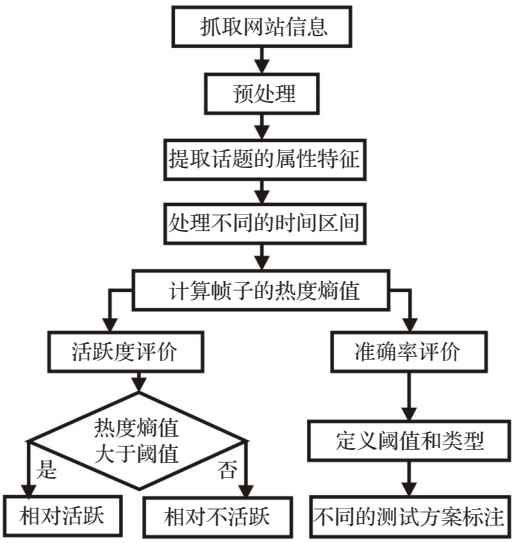

图4　熵值热帖检测流程图

在论坛主题挖掘的过程中，主题特征提取技术将数据中不具备主题特征的数据清除，并能提取出具有特定特征的主题相关数据，主题文本聚类技术将具备主题特征的数据划分为不同的主题集合，而主题热度计算能挖掘出其中的热门主题集合。但在论坛主题挖掘的实际过程，由于挖掘任务的不同，还需通过其他挖掘算法解决不同问题，如对于论坛不同主题的用户情感跟踪[32]，需在主题挖掘的基础上，设计针对于用户的情感挖掘算法，而对于不同主题间演化关系的跟踪，还需设计主题关联算法[33]等等。

3.3　主题建模

主题挖掘算法可获取具有相似主题特征的数据集合，但仍无法自动生成主题。而主题建模可以在保存底层统计关系的同时，为文档语料库提供简洁的主题描述[34]。因此使用主题模型可为各个主题集合生成主题

描述,完成论坛的主题挖掘。

潜在语义索引LSI(latent semantic indexing)[35]是主题模型的雏形,LSI为解决VSM处理结果维数过高、特征稀疏和语义缺失等问题,将文档从高维的单词空间降维映射到低维的潜在语义空间,LSI并非严格意义上的主题模型,但其中潜在语义空间的思想为后续主题模型的研究打下了基础。为改进LSI的不足,Hofmann提出了概率潜在语义分析PLSA(Probability Latent Semantic Analysis)[36],该模型是一个概率生成模型,其设计思想是通过文档层概率、主题层概率和单词层概率来获取与文档最为相关的语义,但存在着过拟合、泛化能力差等问题。随后,David M.Blei在PLSA模型的基础上提出了潜在狄里克雷分布LDA(Latent Dirichlet Allocation)[37],目前论坛主题模型的相关研究大多建立在LDA模型的基础上。

LDA模型包含词、主题、文档三层结构,每个文档由不同的主题按比例混合,同时每个主题都是固定词表上的一个多项式分布。LDA的输出结果是主题-词概率分布的近似值和文档-主题概率分布的近似值,因此该模型能反映词、主题、文档这三个层次之间的关系。值得注意的是,主题的数目需根据具体情况而定。许多研究者在进行论坛主题挖掘探究的过程中,将帖子作为文档,基于对LDA模型的改进,提出了一系列适用于论坛的主题模型。Zhou等[38]开发了基于LDA的交互式在线版——iolda,为输出随时间变化的主题序列,该模型使用滑动窗口技术将论坛文本流划分为一个个时间片,然后对每个时间片使用LDA算法,为发现时间线上的主题演化情况,使用KL相对熵来定量分析主题变异,从而鉴别新主题,并沿着时间线划分出主题序列,但由于iolda在进行时间片划分时,未考虑每个时间片内部的帖子发布时间顺序,且未鉴别不同输入流之间是否存在交集,因此得到的主题序列可能存在顺序误差或主题重叠。主题概率模型所生成的主题出现重叠现象还有一种可能原因是没有选取最佳的主题数目,Jiang等[39]为解决这一问题,在LAD主题模型的基础上引入了主题聚类技术,在使用相似规模的训练集进行聚类实验后,选取最合适的主题噪声阈值和聚类中心数目,然后使用*K*-means++算法对主题空间中的文本集合进行聚类,再通过主题聚簇评价法对各个聚簇出现热门主题的可能性赋予权重,最后从最有可能出现热门主题的聚簇中提取出话题信息,该方案将文本聚类技术与主题概率模型相结合,从文本相似性的角度帮助主题概率模型确定主题数目,从而减少主题重叠。从论坛自身的数据特点考虑,主题存在着明显的依赖性和漂移性,这两种特性是探究论坛主题演化的基础,Ren等[40]提出了一种基于传统的LDA主题模型的回帖传播模型PPM(Post Propagation Model),该模型利用帖子之间的回复关系来动态调整各个主题的分布,将帖子以回复树的形式结构化表示,再将回复树合并为回复图,如图5所示,同时对回复图内部的依赖关系使用LDA主题模型建模,对回复树之间的依赖关系采用Dirichlet分布描述,从而完成主题识别,该模型从论坛结构的角度挖掘主题之间的依赖关系和漂移情况,以此探究论坛的主题演化。同样为探究论坛主题的演化情况,Xu等[41]对论坛文档构建LDA模型后,得到全局主题,然后使用对称的KL divergence算法计算相邻时间段上的主题距离,结合全局主题和各时间段的主题关联,并沿着时间线将关联程度高的主题串联起来,即为主题演化图。Luo等[4]为探索学习论坛的主题与学习者之间的关系,将LDA三层结构中的文档替换为学习者,类似地得到主题-词概率分布的近似值和学习者-主题概率分布的近似值,为确定主题数目,设置跨度为1~100个主题的数据集,进行复杂度评估,选取使模型具有最小复杂度的主题数目,实验结果表明,主题存在重叠现象,可见复杂度评估并不能得到主题概率模型所需的最佳主题数。在使用LDA主题概率模型得到概率近似值后,都需对近似值进行估值,常用的估值方法有Expectation propagation[42],VEM[43]和Gibbs Sampling[44]等。其中Gibbs Sampling算法通过迭代采样来逼近真实的概率分布,实现相对简单,所以应用较为广泛,以上文献均使用了Gibbs Sampling算法。而在主题数目的确定上,各文献使用的方法有所不同,文献[38]通过分析主题变异来确定新主题的数目,文献[40]和文献[41]使用极大似然法来计算主题数目,文献[39]则通过相似度聚类的方法得到最合适的主题中心数,文献[4]针对不同主题数规模进行复杂度评估实验,依据实验结果选取主题数目。以上确认主题数目的方法中,极大似然法是较为通用的方法,而文献[38]

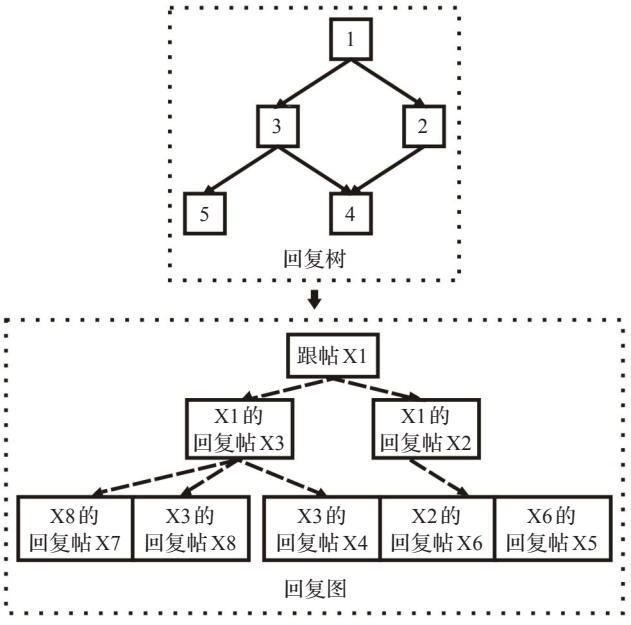

图5 回帖传播模型PPM的回复图

的方法适合于动态追踪主题的演变情况，相反，文献[39]的方法将具有相似度的主题集融合，无法有效识别主题的演变，但能静态识别出不同的主题，文献[4]进行了大量的复杂度评估实验，选取了使模型具有最小复杂度的主题数，但由于过度追求最小复杂度，而导致了主题重叠的发生。

上述LDA主题模型的研究中有个别研究探讨了论坛的主题演化模式，事实上，TDT（Topic Detection and Tracking）是最早用于研究话题演化的方法，它源于1996年美国国防高级研究计划委员会（DARPA）提出需要一种能自动确定新闻报道流中的话题结构的技术，随后，DARPA、卡内基·梅隆大学、Dragon系统公司以及马萨诸塞大学的研究人员开启了对主题演化的研究[45]。上述LDA主题模型中的主题演化研究将论坛数据以时间区间划分，将划分后的子集作为时间序列数据，以此作为主题演化的节点，但实际上区间内部的数据仍是随机无序的，并且时间区间的粒度选取也是误差来源之一。而主题流模型可在一定程度上避免上述问题，与主题模型类似，主题流模型最初的使用对象同样是文档类型数据，此类模型主要探究文档内的主题与该文档的关系，以及文档内各个主题之间的联系，以帮助读者掌握文档的叙述脉络。Blei等[46]提出了动态主题模型DTM（Dynamic Topic Model），实现了时间维度上的文档主题流捕捉。Jeong等[47]提出了实体组主题模型EGTM（Entity Group Topic Model）和连续实体组主题模型S-EGTM（Sequential Entity Group Topic Model），前者可通过主题间的联系获取主题流，后者可通过文档中的片段获取其中各个主题所在的主题流。而在针对论坛的主题挖掘研究工作中，Wu等[48]基于主题流和时间流逝因子，提出了基本主题流模型（B-TFM）和特定主题流模型（T-TFM），并结合主题的随机游动，完成了在线论坛的动态多主题建模，该模型能模拟主题讨论的演化过程并预测用户的参与趋势。其中B-TFM不考虑潜在主题，动态检测论坛的全局主题。T-TFM结合时间流逝因子，识别与特定主题相关的帖子信息。该主题流模型的最大特点在于不必划分时间区间，也能挖掘出论坛主题的演化过程，且可以分别检测出全局主题和特定主题，以满足不同的论坛主题挖掘需求。

总体上来说，主题模型可生成主题描述，主题流模型可获取主题与主题间以及主题与上下文之间的联系。随着论坛主题挖掘研究的不断深入，事实证明，在主题建模中融入除论坛数据以外的其他数据，能生成更具实际价值的内容。例如，He等[5]将学生论坛主题与Guttman量表结合，通过非负矩阵分解NMF（Nonnegtive Matrix Factorization）将Guttman量表纳入到主题模型中，实现了学习效果的评估和学习成果的预测。Atapattu等[49]为帮助MOOCs平台的授课教师在学习讨论区仅查看关于学科的讨论内容，将学科的知识标签与主题模型生成的论坛主题做相似性计算，得到主题标记，有针对性地为教师展现与课程相关的讨论内容。另外，Zhou等[50]为在论坛中找出意见领袖，将帖子间的回复网络结构与主题模型相结合，从论坛数据中识别出具有较高实用性主题属性的意见领袖。以上均是主题建模在实际问题中的有效运用。

4　存在问题与研究展望

4.1　存在问题

国内外学者在论坛主题挖掘方面上取得了诸多成果，但这些研究现状仍存在一定的问题，具体问题如下：

4.1.1　代词导致主题信息缺失

论坛用户以主题词汇为核心展开讨论，但往往在对话中主题词汇会被代词代替，这将大大影响真实主题词汇的识别频率，导致主题识别发生偏差，且随着代词使用率的增加，偏差会越发明显。解决这一问题，即还原真实主题词汇的出现频率，需要识别代词的实际含义。

对代词实际含义的识别离不开句子间的语义关系，但由于论坛句子的上下文并不具有相邻位置关系，而是具有回复关系，因此需结合线程的回复结构来识别论坛中代词的含义，但目前关于论坛线程结构的研究还不够成熟。

4.1.2　时间区间的划分

论坛主题演化的研究都需要加入时间要素，以此跟踪主题在时间线上的变化情况，但有诸多研究的做法是将论坛文本按一定的粒度划分为时间区间，并使用主题挖掘技术处理各时间区间的文本。但事实上，各时间区间内部的数据仍是随机无序的，并且时间区间粒度的选取过大或过小也将导致误差，粒度过大会使区间内的主题演化无法被检测到，而粒度过小会使主题偏移不断累积，甚至脱离主题。

4.2　研究展望

论坛主题挖掘是一个结合了多学科的研究问题，它涉及语义分析、自然语言处理、概率统计、聚类、机器学习等领域，随着相关技术的发展，未来对论坛主题的研究可集中在以下几个方面。

4.2.1　论坛热点话题预测

随着网络论坛用户数量的增长和用户线上活跃度的提升，网络论坛的社会影响力越来越大，如今许多社会热点新闻都最先出自于各网络论坛，随后被广泛讨论并传播。因此，如果能够预测论坛热点话题，或在检测论坛热点话题的基础上预测其发展趋势，对于政府的舆情调控和商业趋势的发现等都有着重要意义。

对话题预测技术的研究广泛应用于舆情监测领域。近年来，关于话题预测技术的研究有很多，其中许多研究以隐马尔科夫模型[51-52]和神经网络[53]等传统预测

模型为基础,除此之外,还有学者提出了基于自适应AR模型[54]、基于灰色Verhulst模型[55]和基于小波变换[56]的主题预测算法。

以上研究为论坛的热点话题预测奠定了基础,但此方向仅处于起步阶段,目前关于论坛热点话题预测的研究还比较少,包括Xu等[57]基于BP神经网络提出了一种论坛话题的热度预测方法,Cheng等[58]基于人工神经网络提出了一种预测在线论坛话题增长趋势的方法。

4.2.2 基于主题挖掘的论坛检索功能

由于论坛的特殊结构,论坛用户往往难以快速从大量的帖子中找到所需内容,而目前实际应用于论坛的检索功能大多基于关键词与主帖标题的匹配,且有关于论坛检索的科学研究主要集中在将帖子的哪些特征作为检索结果的排序依据[59-60],或将论坛的线程结构抽象化,从中提取了最佳问答对话,以此构建论坛的检索系统[61-62],但尚未有研究关注到跟帖中的讨论主题其实无法被检索到的问题。

如果将论坛主题挖掘结果中的各个主题与该主题的源文本相匹配,并提取各主题中的所有有效关键词,以此应用于论坛检索中,将能有效解决这一问题。

4.2.3 论坛主题可视化

随着可视化技术的发展,将各种复杂多元甚至实时数据精确地进行图像显示已不再是难事。对于论坛用户或论坛管理者而言,为他们提供直观的论坛主题可视化图像能帮助他们高效掌握论坛的主要讨论内容。

目前存在许多关于主题可视化的研究,如Liu等[63]实现了文本流中各主题文本摘要的可视化显示,Cao等[64]使用太阳映射技术展示了各主题间的层次关系,Gretarsson等[65]实现了大型文本语料库的全局主题可视化并允许查看每个主题下的子题及其文本内容。

而对比现有的论坛主题可视化案例,如Speck等设计的论坛可视化系统ForumDash[66]和Qu等设计的论坛可视化分析系统iforum[67],其中对论坛主题的可视化实现较为单一,因此将论坛主题挖掘过程与可视化技术有效结合仍具有较大的发展空间。

5 结束语

论坛主题挖掘研究的关键是从论坛文本流中识别出用户集中讨论的内容,并从中提取主题。本文首先对论坛主题挖掘研究进行了问题描述和框架梳理,然后对现有技术进行了分类,并详细阐述了各类论坛主题挖掘技术,最后对当前存在的问题和发展趋势做了比较详细的分析和讨论。该研究领域仍存在着一定的的问题和挑战,深入的研究将能进一步提高论坛数据的应用价值。

参考文献:

[1] Zhang Y,Zhang H.Social topic detection for web forum[C]// International Conference on Computer Science & Service System,Washington,DC,USA,August 11-13,2012.Washington,DC,USA:IEEE Computer Society,2012:955-959.

[2] Chen Y,Yang S,Cheng X Q.Bursty topics extraction for web forums[C]//Proceedings of the Eleventh International Workshop on Web Information and Data Management,Hong Kong China,November 02-02,2009.New York,USA:ACM,2009:55-58.

[3] MA Xiu-Lin,JIN Hai-yan.The research about the organization of the instruction forum topic by keyword marking[J].Modern Educational Technology,2009,12.

[4] Luo C,He T,Zhang X,et al.Learning forum posts topic discovery and its application in recommendation system[J].Journal of Software,2015,10:392-402.

[5] He J,Rubinstein B I P,Bailey J,et al.MOOCs meet measurement theory:a topic-modelling approach[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence,Phoenix,Arizona,February 12-17,2016.[S.l.]:AAAI Press,2016:1195-1201.

[6] 化柏林.知识抽取中的停用词处理技术[J].现代图书情报技术,2007,2(8):48-51.

[7] 刘群,张华平,俞鸿魁,等.基于层叠隐马模型的汉语词法分析[J].计算机研究与发展,2004(8):1421-1429.

[8] Qiu X,Zhang Q,Huang X.FudanNLP:a toolkit for Chinese natural language processing[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics:System Demonstrations,2013:49-54.

[9] Tseng H,Chang P,Andrew G,et al.A conditional random field word segmenter[J].Foundations of Science,2005:168-171.

[10] Zhao H,Huang C N,Li M,et al.An improved Chinese word segmentation system with conditional random field[C]//Proceedings of the Fifth Sighan Workshop on Chinese Language Processing,2006:162-165.

[11] Li X,Dai G,Lai S,et al.Hot topic detection in Chinese web forum using statistics approach[C]//IEEE International Conference on Signal Processing,Communications and Computing,Xi'an,China,September 14-16,2011.IEEE,2011:1-4.

[12] Salton G,Wong A,Yang C S.A vector space model for automatic indexing[J].Communications of the ACM,1975,18(11):613-620.

[13] Yan T,Wang X W.Feature extension for short text[C]//Proceedings of the International Symposium on Computer Science,2010.

[14] He H,Chen B,Xu W,et al.Short text feature extraction and clustering for web topic mining[C]//Third International Conference on Knowledge Grid and Semantics,Shanxi,China,October 29-31,2007.IEEE,2007:382-385.

[15] Li S,Lv X,Li Y,et al.Study on feature selection algorithm in topic tracking[C]//2010 2nd International Conference on Software Engineering and Data Mining (SEDM),Chengdu,China,June 23-25,2010.IEEE,2010:384-389.

[16] 高飞,周学广,孙艳.基于关联规则和粗糙集的话题特征提取方法[J].计算机工程,2012,38(10):63-66.

[17] Chen Y,Cheng X Q,Huang Y L.A wavelet-based model to recognize high-quality topics on web forum[C]//IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology,Sydney,NSW,Australia,December 9-12,2008.IEEE,2008:343-351.

[18] Chen F,Du J,Qian W,et al.Topic detection over online forum[C]//Web Information Systems and Applications Conference,Haikou,China,November 16-18,2012.IEEE,2012:235-240.

[19] Wang L,Dai G Z.Forum hot topic detection based on community structure of complex networks[J].Computer Engineering,2008,34(11):214-217.

[20] Zhang L.Prediction of the attention of internet forum hot topics[J].Computer & Digital Engineering,2013(5).

[21] Hartigan J A,Wong M A.Algorithm AS 136:A *k*-means clustering algorithm[J].Journal of the Royal Statistical Society Series C:Applied Statistics,1979,28(1):100-108.

[22] Mclachlan G J,Krishnan T.The EM algorithm and extensions[J].Biometrics,1997,382(1):154-156.

[23] Ester M,Kriegel H P,Sander J,et al.A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise[C]//International Conference on Knowledge Discovery and Data Mining,Portland,Oregon,August 02-04,1996.[S.l.]:AAAI Press,1996:226-231.

[24] Zhang T,Ramakrishnan R,Livny M.BIRCH:an efficient data clustering method for very large databases[J].Acm Sigmod Record,1996,25(2):103-114.

[25] Guha S,Rastogi R,Shim K.Cure:an efficient clustering algorithm for large databases[J].Information Systems,2001,26(1):35-58.

[26] Ng R T,Han J.CLARANS:A method for clustering objects for spatial data mining[J].IEEE Transactions on Knowledge & Data Engineering,2002,14(5):1003-1016.

[27] 梅泽勇,王清飞.基于BBS的热点问题发现[J].情报探索,2011(3):14-16.

[28] Sun Y L,Dong L I,Zhang Y.Hot topics foundation in network forum based on entropy[J].Computer Engineering,2014,40(6):312-316.

[29] 谌志群,徐宁,王荣波.基于主题演化图的网络论坛热点跟踪[J].情报科学,2013(3):147-150.

[30] 陈麓屹,周斌彬,徐萍.虚拟社区话题热度评估算法研究[J].浙江树人大学学报:自然科学版,2015(1):1-4.

[31] 杜慧,郭岩,范意兴,等.基于因果模型的主题热度计算与预测方法[J].中文信息学报,2016,30(2):50-55.

[32] Ramesh A,Kumar S H,Foulds J R,et al.Weakly supervised models of aspect-sentiment for online course discussion forums[C]//Annual Meeting of the Association for Computational Linguistics,Beijing,China,July 26-31,2015.ACL:2015:74-83.

[33] Lavrenko V,Allan J,Deguzman E,et al.Relevance models for topic detection and tracking[C]//HLT'02 Proceedings of the Second International Conference on Human Language Technology Research,San Diego,California,March 24-27,2002.San Francisco,CA,USA:Morgan Kaufmann Publishers Inc,2002:115-121.

[34] Wu H,Bu J,Chen C,et al.Locally discriminative topic modeling[J].Pattern Recognition,2012,45(1):617-625.

[35] Deerwester S,Dumais S T,Furnas G W,et al.Richard harshman indexing by latent semantic analysis[J].Journal of the American Society for Information Science,1990.

[36] Hofmann T.Probabilistic latent semantic indexing[C]//International ACM SIGIR Conference on Research and Development in Information Retrieval,Berkeley,California,USA,August 15-19,1999.New York,NY,USA:ACM,1999:50-57.

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

[38] Zhou B,Kai C,Jia Y,et al.Interactive mining topic evolutionary patterns from internet forums[C]//International Conference on Education Technology and Computer,Shanghai,China,June 22-24,2010.IEEE,2010,5:76-81.

[39] Jiang H,Chen XF,Du M.On-line forum hot topic mining method based on topic cluster evaluation[J].Journal of Computer Applications,2013,33(11):3071-3075.

[40] Ren Z,Ma J,Zhumin A C.Web forum thread summarization based on dynamic topic modeling[J].Journal of Computer Research & Development,2012,49(11):2359-2367.

[41] 徐佳俊,杨飏,姚天昉,等.基于LDA模型的论坛热点话题识别和追踪[J].中文信息学报,2016(1):43-49.

[42] Takita M,Naziruddin B,Matsumoto S,et al.Expectation-propogation for the generative aspect model[J].Computer Science,2012,235(11):3257-3269.

[43] Steyvers M,Griffiths T.Handbook of latent semantic analysis[J].Wiley Interdisciplinary Reviews Cognitive Science,2007,4(1):683-692.

[44] Ishwaran H,James L F.Gibbs sampling methods for stick-breaking priors[J].Journal of the American Statisti-

cal Association,2001,96(453):161-173.

[45] 曹丽娜,唐锡晋.基于主题模型的BBS话题演化趋势分析[J].管理科学学报,2014,17(11):109-121.

[46] Blei D M,Lafferty J D.Dynamic topic models[C]//ICML'06 Proceedings of the 23rd International Conference on Machine Learning,Pittsburgh,Pennsylvania,USA,June 25-29,2006.New York,NY,USA:ACM,2006:113-120.

[47] Jeong Y S,Choi H J.Sequential entity group topic model for getting topic flows of entity groups within one document[M]//Tan P N,Chawla S,Ho C K,ed.Advances in Knowledge Discovery and Data Mining.Berlin Heidelberg:Springer,2012,7301:366-378.

[48] Wu H,Bu J,Chen C,et al.Modeling dynamic multi-topic discussions in online forums[C]//AAAI'10 Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence,Atlanta,Georgia,July 11-15,2010.[S.l.]:AAAI Press,2010:1455-1460.

[49] Atapattu T,Falkner K.A framework for topic generation and labeling from MOOC discussions[C]//Proceedings of the Third (2016) ACM Conference on Learning,Edinburgh,Scotland,UK,April 25-26,2016.New York,NY,USA:ACM,2016:201-204.

[50] Zhou X,Yang J,Zhang J,et al.A BBS opinion leader mining algorithm based on topic model[J].Journal of Computational Information Systems,2014,10(6):2571-2578.

[51] Liu R,Guo W.HMM-based state prediction for Internet hot topic[C]//IEEE International Conference on Computer Science and Automation Engineering,Shanghai,China,June 10-12,2011.IEEE,2011:157-161.

[52] Zeng J,Zhang S,Wu C,et al.Predictive model for internet public opinion[C]//International Conference on Fuzzy Systems and Knowledge Discovery,Haikou,China,August 24-27,2007.IEEE Computer Society,2007:7-11.

[53] Lu H Y,Xie L Y,Kang N,et al.Don't forget the quantifiable relationship between words:using recurrent neural network for short text topic discovery[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence,San Francisco,California,USA,February 4-9,2017.[S.l.]:AAAI Press,2017:1192-1198.

[54] Tong H,Liu Y,Peng H,et al.Internet users' psychosocial attention prediction:web hot topic prediction based on adaptive AR model[C]//International Conference on Computer Science and Information Technology,Singapore,August 29-September 2,2008.IEEE,2008:458-462.

[55] Wang X,Qi L,Chen C,et al.Grey System Theory based prediction for topic trend on Internet[J].Engineering Applications of Artificial Intelligence,2014,29(3):191-200.

[56] Fang M,Chen Y,Gao P,et al.Topic trend prediction based on wavelet transformation[C]//Web Information System and Application Conference,Tianjin,China,September 12-14,2014.IEEE,2015:157-162.

[57] Xu T,Xu M,Ding H.BBS Topic's hotness forecast based on back-propagation neural network[C]//International Conference on Web Information Systems and Mining.IEEE Computer Society,Sanya,China,October 23-24,2010.IEEE,2010:57-61.

[58] Cheng J J,Liu Y,Cheng H,et al.Growth trends prediction of online forum topics based on Artificial neural networks[J].Journal of Convergence Information Technology,2011,6(10):87-95.

[59] 杨小锐,林磊,孙承杰,等.基于结构挖掘的论坛检索模型[J].中文信息学报,2011,25(1):98-104.

[60] Huang Y M,Chen J N,Kuo Y H,et al.An intelligent human-expert forum system based on fuzzy information retrieval technique[J].Expert Systems with Applications,2008,34(1):446-458.

[61] Singh A,Deepak P,Raghu D.Retrieving similar discussion forum threads:a structure based approach[C]//SIGIR'12 Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval,Portland,Oregon,USA,August 12-16,2012.New York,NY,USA:ACM,2012:135-144.

[62] Wang L,Kim S N,Baldwin T.The utility of discourse structure in forum thread retrieval[C]//Asia Information Retrieval Symposium,Singapore,December 9-11,2013.Berlin Heidelberg:Springer,2013:284-295.

[63] Liu S,Zhou M X,Pan S,et al.Interactive,topic-based visual text summarization and analysis[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management,Hong Kong,China,November 02-06,2009.New York,NY,USA:ACM,2009:543-552.

[64] Cao N,Gotz D,Sun J,et al.SolarMap:multifaceted visual analytics for topic exploration[C]//2011 IEEE 11th International Conference on Data Mining (ICDM),Vancouver,BC,Canada,December 11-14,2011.IEEE,2011:101-110.

[65] Gretarsson B,Donovan J,Bostandjiev S,et al.TopicNets:visual analysis of large text corpora with topic modeling[J].ACM Transactions on Intelligent Systems & Technology,2012,3(2).

[66] Speck J,Gualtieri E,Naik G,et al.ForumDash:analyzing online discussion forums[C]//ACM Conference on Learning,2014:139-140.

[67] Fu S,Zhao J,Cui W,H Qu.Visual analysis of MOOC forums with iForum[J].IEEE Transactions on Visualization & Computer Graphics,2016,23(1):201-210.