软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(2):246−261 [doi: 10.13328/j.cnki.jos.005116]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于动态主题模型融合多维数据的微博社区发现算法*

刘冰玉[1], 王翠荣[1], 王 聪[1], 王军伟[1], 王兴伟[2], 黄 敏[1]

[1](东北大学 计算机科学与工程学院,辽宁 沈阳 110819)

[2](东北大学 软件学院,辽宁 沈阳 110819)

通信作者: 王兴伟, E-mail: wangxw@mail.neu.edu.cn

摘 要: 随着微博用户的不断增加,微博网络已成为用户进行信息交流的平台.针对由于博文长度受限,传统的社区发现算法无法有效解决微博网络的稀疏性等问题,提出了 DC-DTM(discovery community by dynamic topic model)算法.DC-DTM 算法首先将微博网络映射为有向加权网络,网络中边的方向反映节点之间的关注关系,利用所提出的 DTM(dynamic topic model)计算出节点之间的语义相似度,并将其作为节点间连边的权重.DTM 是一种微博主题模型.该模型不仅能够挖掘博客的主题分布,而且能够计算出某一主题中用户的影响力大小.其次,利用所提出的复杂度较低的标签传播算法 WLPA(weighted lebel propagation)进行微博网络的社区发现.该算法的初始化阶段将影响力大的用户节点作为初始节点,标签按照节点的影响力从大到小进行传播,避免了传统标签传播算法逆流现象的发生,提高了标签传播算法的稳定性.真实数据上的实验结果表明,DTM 模型能够很好地对微博进行主题挖掘,DC-DTM 算法能够有效地挖掘出微博网络的社区.

关键词: 新浪微博;文本挖掘;DC-DTM;吉布斯采样;LDA;主题模型

中图法分类号: TP181

中文引用格式: 刘冰玉,王翠荣,王聪,王军伟,王兴伟,黄敏.基于动态主题模型融合多维数据的微博社区发现算法.软件学报, 2017,28(2):246−261. http://www.jos.org.cn/1000-9825/5116.htm

英文引用格式: Liu BY, Wang CR, Wang C, Wang JW, Wang XW, Huang M. Microblog community discovery algorithm based on dynamic topic model with multidimensional data fusion. Ruan Jian Xue Bao/Journal of Software, 2017,28(8):246−261 (in Chinese). http://www.jos.org.cn/1000-9825/5116.htm

Microblog Community Discovery Algorithm Based on Dynamic Topic Model with Multidimensional Data Fusion

LIU Bing-Yu[1], WANG Cui-Rong[1], WANG Cong[1], WANG Jun-Wei[1], WANG Xing-Wei[2], HUANG Min[1]

[1](School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China)

[2](School of Software, Northeastern University, Shenyang 110819, China)

Abstract: With the dramatic increase of microblog users, microblog websites have become the platform for a wide spectrum of users to get information. Due to the fact that blog is a special type of text with restricted length, traditional community detection algorithms cannot effectively solve the sparse problem of micro blog. To address the issue, the DC-DTM (discovery community by dynamic topic model)

* 基金项目: 国家杰出青年科学基金(61225012, 71325002); 国家自然科学基金(61572123, 61300195); 高等学校博士学科点专项科研基金(20120042130003); 辽宁省百千万人才工程项目(2013921068); 河北省自然科学基金(F2014501078); 河北省科技计划(15210146)

Foundation item: National Science Foundation for Distinguished Young Scholars of China (61225012, 71325002); National Natural Science Foundation of China (61572123, 61300195); Specialized Research Fund of the Doctoral Program of Higher Education (20120042 130003); Liaoning BaiQianWan Talents Program (2013921068); Natural Science Foundation of Hebei Province (F2014501078); Technology Planning Project of Hebei Province (15210146)

收稿时间: 2015-12-26; 修改时间: 2016-03-17; 采用时间: 2016-06-10; jos 在线出版时间: 2016-10-11

CNKI 网络优先出版: 2016-10-12 16:27:00, http://www.cnki.net/kcms/detail/11.2560.TP.20161012.1627.027.html

algorithm is proposed in this paper. First, the algorithm maps microblog as a directed-weighted network, in which the direction is the concerned relationship, and the weight is the topic's similarity of different nodes calculated by DTM (dynamic topic model). DTM is a microblog topic model which can not only mine the topics of each microblog accurately but also calculate author's influence a topic. Second, the algorithm uses label propagation WLPA (weighted lebel propagation), with low complexity, to find communities in microblog. The initial process selects nodes with the largest influence as the initial nodes, and propagates the label in the order of node's influences, from large to small. The algorithm overcomes the adverse phenomenon in the traditional label propagation algorithm, and has better stability. Experiments on real data show that the DTM model can be very good for the topic mining in microblog and DC-DTM algorithm can effectively discover the communities of microblog.

Key words: Sina microblog; text mining; DC-DTM; Gibbs sampling; LDA; topic model

现实生活中存在多种类型的复杂网络,包括生物网络、文献引文网络以及新兴的社会网络.复杂网络具有小世界、无标度等特点,对复杂网络的结构特征分析已经成为当前研究的热点.社交网络是一种新兴的复杂网络,遵循复杂网络的特点,对社交网络的社区划分能够为商业推荐和用户管理以及研究工作者提供有意义的帮助.

微博(microblog)是 Web 2.0 时代的标志性产物,是一种集成化、开放性的社交网络,是基于用户关系的信息发布、传播以及获取的平台.作为一种新兴的社交网络平台,用户创造的不同格式(文本、图片)的数据传播到微博网络中,不仅受到了用户的欢迎,而且也获得了很多研究人员的关注,已经成为国内外专家的研究热点.在微博网络中,用户可以发布简短文本(通常少于 140 字)信息,此类信息称为博客(或博文),而博文信息是通过用户之间的转发或评论等方式在网络中进行传播.作为新兴媒体,微博网络具有简便、及时等特点,受到网络用户的日益青睐,信息涵盖的主题比传统媒体更加宽泛,有很高的更新与传播速度.对微博网络的用户关系进行分析,并对潜在主题进行挖掘,找出兴趣相似的用户以及特定领域的活跃用户,掌握社区动态,可以为舆情控制及个性化推荐等提供有力的支撑.

微博网络中的社区是由具有相同或相似兴趣爱好的用户组成的,可以分享并交流信息的子团体.由于微博网络具有用户规模大、节点的度分布不均匀、弱社交网络等特点,传统的社区发现算法应用于微博网络具有局限性,而基于概率主题模型的社区发现算法不仅适用于大规模网络,而且能够将微博网络中节点的语义信息和微博网络拓扑结构融合,主要的概率主题模型有 LDA(latent Dirichlet allocation)[1],ATM(author topic model)[2]等.

由于博文为短文本,存在数据稀疏性问题,本文提出 DTM(dynamic topic model)主题模型来挖掘微博网络中的作者博文.考虑到博文的动态性,只选取作者近一年的博文作为研究对象,将不同作者博文的主题相似度作为作者之间链接的权重,将微博网络映射为有向加权网络,并提出复杂度低的标签传播算法 WLPA(weighted lebel propagation)进行社区发现.通过在新浪微博数据上进行实验,结果表明:本文所提出的主题模型 DTM 与 LDA 模型相比,模型的困惑度较低,并且该模型能够较为准确地定位对某一话题敏感的作者集合;本文所提出的 DC-DTM 社区发现算法发现的社区的模块度较高.

本文的主要贡献如下:(1) 根据微博网络中节点自身以及节点之间联系的特性,建立微博网络的 RT 模型,将微博网络映射为有向加权网络;(2) 利用 DTM 概率模型,将转发/评论博文与原博文相结合,挖掘作者的主题分布以及整个网络的主题分布趋势;(3) 将不同作者的主题分布相似度作为作者之间联系的权重,在此基础上利用标签传播算法对微博网络进行重叠社区发现;(4) 针对微博网络的社区发现,提出了基于语义的微博网络社区的评价指标;(5) 在大量新浪微博数据中进行实验,实验结果表明本文所提出的 DTM 模型在生成主题质量、内容困惑度以及模型复杂度等方面的指标都优于传统 LDA 模型,本文的 DC-DTM 算法发现的社区比传统的基于标签的社区发现算法的模块度高,稳定性好.

1 相关工作

目前研究者已经提出了许多社区发现算法,从传统的根据网络拓扑结构进行社区发现发展到基于语义的社区发现以及融合拓扑结构和语义的社区发现算法.下面分别对这 3 类算法进行阐述.

1.1 基于网络拓扑结构的算法

微博网络中用户之间的关注或好友关系形成了用户网络结构,此类方法以图论为基础进行社区挖掘,主要分为非重叠社区划分和重叠社区划分.

所谓非重叠社区划分,即此类方法所划分的社区中,某个节点只能属于某一个社区,不能同时属于多个社区,此类方法是最早出现的社区划分方法.算法主要有谱方法、模块度优化方法、层次聚类方法、边预测方法等[3−6].

现实生活中,一个节点可能以不同的角色参与到不同的社区,重叠社区结构更贴近网络的真实结构.2005年,Palla 等人[7]对传统的社区发现方法进行扩展,允许一个节点同时隶属于多个不同的社区,提出了具有重叠特性的社区结构,并设计了重叠社区发现算法.在重叠社区发现中,重叠节点是社区之间的桥梁,对社区演化和社区间互通信息起到关键作用.当社区间的重叠度高时,表明共同拥有大部分的成员,那么这些社区即将融合为一个大社区;反之,若社区间的重叠度低,则说明网络的社区间相对独立.因此对社区重叠结构的研究不仅更贴近网络实际的社区结构组成,而且也有助于分析网络的演化趋势.此类算法主要是 CPM(clique percolation method)算法[7].该算法的基本思想是社区内部的节点由于具有较高的密度而更容易形成派系;分区的优化算法都是对社区的基本度量标准,如密度、模块度和导电率等进行优化.其中,OSLOM(order statistics local optimization method)[8]算法是通过增加或删除社区内的节点来增强社区的适应度函数;最著名的是,Baumes 等人在文献[9]中提出了一种两阶段方法,首先将网络划分为不相交的“种子”社区,然后通过增加或移除某些邻接节点来达到社区的“密度”最大的效果.这些算法大部分都存在需要先验知识或时间复杂度较高的缺点.

由于社会网络的数据集非常庞大,社区发现算法的速度变得尤为重要.迄今为止速度最快的算法之一就是 Raghavan 等人[10]提出的 RAK 算法.Raghavan 等人的社区发现算法除了具有线性时间复杂度的特性以外还具有简单和无需先验知识等优点.但是,像大多数社区发现算法一样,它只能发现不重叠的社区.由于标签更新阶段是按照随机的顺序对节点进行更新,并未考虑到节点在网络中的重要程度,这会导致标签传播过程的“逆流”现象,影响算法的稳定性.

研究人员围绕 Raghavan 等人提出的标签传播算法进行改进,分别在算法的传播规则、收敛条件和更新策略等方面进行相关研究.但这些算法都只能发现非重叠社区.随后研究人员将标签传播算法进一步改进,使其应用于重叠社区的发现.第一个使用标签传播算法解决重叠社区发现问题的是 Steve 提出的 COPRA 算法(community overlap propagation algorithm)[11].该算法虽然能够解决重叠社区的发现问题,但 COPRA 算法要求输入节点所属重叠社区的数量,这对于大型网络并不适用.在大型网络中,为网络设置一个全局的重叠社区个数上限很难符合不同网络、不同节点所处的网络情况,并且算法仍然存在不稳定等缺点.还有一些算法针对标签传播算法的稳定性进行改进,但这些算法都存在一定的局限性.文献[12]提出了一种基于标签传播概率的 LPPB(label-propagation-probability-based)重叠社区发现算法.该算法在标签传播的过程中,综合网络的结构传播特性和节点的属性特征共同计算标签传播的概率,同时利用节点的历史标签记录修正标签更新结果;最后将传播后具有相同标签的节点划分为同一社区,社区间的重叠节点构成了社区重叠结构.

此类方法仅仅考虑了用户之间的关系,并未考虑用户的内容,而微博网络中用户的内容能够体现用户的兴趣,对于衡量用户之间的相似度具有指导意义.

1.2 基于用户内容的算法

考虑到微博网络中除了用户之间的关注关系之外,用户发布的博文是用户兴趣的真实反映,是用户兴趣的重要依据,以用户内容为研究对象的微博网络社区发现算法应运而生.此类方法主要通过计算博文内容的相似性进行文本聚类,将内容相似的博文划分为社区.主题模型是最典型的文本聚类算法.2003 年,Blei 等人引入了 Dirichlet 先验分布,形成了一个“文档-主题-词”3 层的贝叶斯模型 LDA[1].该模型首先提出了“主题”的概念,在主题模型的框架下,主题是语料集合上语义的抽象,是语料集合依赖的,不同的语料潜在的语义是不同的;随后通过概率方法对模型推导,即可挖掘出文本集的语义结构.标准 LDA 模型认为一篇文章的每个词都是通过“以一定概率选择了某个主题,并从这个主题中以一定概率选择某个词语”这样一个过程得到的.

LDA 首先从先验 Dirichlet 分布中抽取该文档的主题分布,然后,从主题多项式分布中选择当前单词的主题;最后,从先验 Dirichlet 分布中抽取该主题的单词分布,并选择具体单词.LDA 模型的参数估计主要采用近似推理算法,方法主要有变分最大期望算法[13]、期望传播算法[14]、吉布斯采样算法[15].其中,吉布斯采样算法由于简单易懂和运行速度快等优点,常被用来求解 LDA 模型.

近年来,随着主题模型研究的深入,研究者开发了多个 LDA 派生模型.这些模型从不同的角度增加辅助信息以解决数据稀疏问题,提高了微博的主题挖掘效果.算法主要有:文献[16]针对微博网络高维稀疏空间的问题提出了基于 LDA 的文本分类方法;文献[17]通过用户标签的使用频率等信息对用户兴趣进行聚类分析,将兴趣相似的用户加入到同一个社区,该方法忽略了微博网络中的关注关系,而微博网络的关注关系恰恰是微博网络信息传播的桥梁,在微博网络中能够及时反映用户的兴趣.文献[18−22]利用 LDA 模型对语义网络进行主题挖掘建立语义空间,分别利用标签传播算法、BLOCK 场、随机游走以及话题综合因子分解等方法进行社区发现.

以上算法在对微博网络进行主题挖掘时,并没有将转发博文或评论博文与博客原文建立联系,也未区分博文的类型,导致算法效果并不理想.

1.3 融合网络结构和用户内容的算法

针对以上两种社区发现算法的不足,研究人员将上述两种方法相结合,根据网络结构提取用户联系,根据用户的博文信息提取用户的兴趣,将用户联系和用户兴趣进行融合实现社区的划分[23−29].代表性的文献主要有:文献[23]构建了以关注关系为网络节点,以关注关系之间是否有共同的用户作为其潜在的边构建了微博网络的 R-C 模型;文献[24]通过融合微博内容和微博之间的交互关系对用户的兴趣进行挖掘;文献[25]将用户和文本内容相融合来发现社区,文章采用 AT(author topic)模型进行用户兴趣社区的发现,采用 NMF 方法进行用户关系社区发现,算法最后将两种社区进行融合来发现最终的社区;文献[26]在研究 LDA 的基础上,对微博的联系人关联关系以及文本之间的关联关系进行统一建模.随着微博的普及,微博数据量急速增长,微博中用户的行为更加丰富,用户的转发、评论等行为是信息在微博网络中传输的原动力,也是用户兴趣的真实反映,而上述模型并未考虑到用户行为的因素,也未对特定主题下用户的影响力进行研究.

以上算法没有充分考虑微博网络的特点,只是将微博网络抽象为无向无权网络或有向无权网络,并未抽象为更贴近现实的有向加权网络.除此之外,以上算法并未对微博网络中的博文进行分类,未将转发或评论博文与相关原创博文进行联系.本文所提出的 DTM(data topic model)模型将评论或转发博文与原博文相关联,不仅解决了数据稀疏性问题,而且能够更好地理解博文内容.除此之外,考虑到博文的动态性以及作者的兴趣变化,该模型主要挖掘作者最近 1 年发布的博文主题,并计算不同作者的博文之间的相似度.由于要将微博网络抽象为有向网络,本文采用 KL(kullback Leibler)距离进行相似度度量,将不同作者之间博文的 KL 距离作为作者的相似度.将作者兴趣相似度作为网络中链接的权重,对微博网络进行建模,将微博抽象成有向加权网络.在此基础上,通过标签传播算法 WLPA 对抽象模型进行社区划分.本文提出的稳定的标签传播算法 WLPA(weighted label propagation algorithm),对关注某一特定主题的用户按照用户的影响力进行排序,在标签传播过程中,标签按照用户的影响力大小进行传播,在增强算法的稳定性、准确性及鲁棒性的同时,保证了算法的线性复杂度的优点.

2 基于 DTM 模型的微博社区挖掘算法 DC-DTM

本文在分析微博网络特性的基础上提出了 DC-DTM 算法.该算法首先建立微博网络的 RT(reply and forward post)模型,将微博网络映射为以博主为节点,以博主之间的关注关系为边,将博文主题之间相似度作为边权重的有向加权网络,其中,博文主题之间的相似度是通过本文所提出的概率模型 DTM 对微博网络中各节点的主题进行挖掘,将不同作者微博的 KL 距离作为其兴趣相似度,即连边的权重.利用该模型挖掘出每个主题中各节点的影响力,将影响力最高的节点作为之后标签传播算法的初始节点,通过标签传播算法 WLPA,对微博网络进行社区发现.

2.1 微博网络 RT 模型

本文首先对微博网络进行抽象建模,由于微博网络中的关注关系具有方向性,并且不同用户之间的联系紧密程度不同,本文将微博网络抽象为有向加权网络.网络中的节点为微博中的用户,如果用户 A 关注用户 B,则存在一条边 E_{AB} 由用户 A 指向用户 B.将用户之间的兴趣相似度作为边的权重.通过对微博网络的研究发现,在微博网络中,以下因素能够阐释用户的兴趣特征.

(1) 发表博文内容

用户只发布自己感兴趣的内容,所以用户以往发布的博文能够直接表征用户的兴趣.

(2) 转发和评论博文的内容

用户在阅读其他用户发布的博文之后,如果对该博文感兴趣,一般会对其进行转发或评论,这种转发或评论的行为也能够映射出作者的兴趣点.本文中将转发和评论行为归为一类,统称为关注.

(3) 用户的标签

用户标签是用户在完善个人资料时指定的一组描述用户兴趣、爱好的关键字,是表征用户兴趣、爱好的一种手段.但用户的兴趣会随时间的推移变化,并且用户一般不会经常更改自己的标签,标签信息对本文研究贡献度较低,本文中不考虑此类内容.

综上所述,本文中将用户的发布或转发/评论的博文内容作为其兴趣特征.利用主题模型对用户的博文内容进行主题挖掘,将作者关注主题之间的相似度作为作者联系的权重.

本文所建立的微博 RT 模型如图 1 所示.在图 1 中,有两个用户发布的两个博文,其中用户 B 发布 4 号博文,用户 C 发布 2 号博文,而用户 A 的 1 号博文关注了用户 B 的 4 号博文,用户 A 的 5 号博文关注了用户 C 的 2 号博文,用户 C 的 3 号博文关注了用户 B 的 4 号博文.由博文之间的关注关系得到用户之间的关联关系,用户 A 关注用户 B 和 C,用户 C 关注用户 B.其中,用户之间的关系为有向边,边的权重为用户兴趣的相似度.例如,E_{AB} 边的权重为用户 A 所发博文 1 和博文 5 与用户 B 所发博文 4 之间的相似度.在此模型中,用户博文的相似度通过博文之间的主题相似度 KL 距离表征.

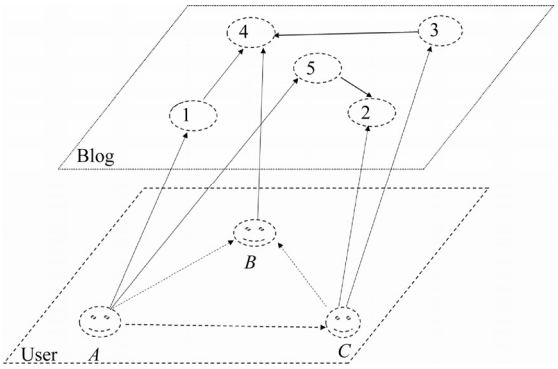

Fig.1 RT model

图 1 RT 模型

至此,在 RT 模型中,微博网络用图模型 $G=(V,E)$表示,其中,V 表示在微博网络中所有节点的集合,每个节点都有内容集合 M_i,表示用户 i 所发布的所有博文(包括转发博文、评论博文),E 表示在节点之间有转发关系的边的集合.

2.2 模型介绍

2.2.1 DTM 模型

针对现有模型的不足,尤其是微博网络的稀疏性问题,在综合考虑微博的时间信息、主题信息、微博的标

签以及作者的兴趣等问题的基础上,本文提出了适用于微博网络的主题模型,对博主的博文进行主题提取.DTM模型是在LDA的基础上,针对微博类网络结构中潜在的作者联系信息,集成主题挖掘、作者兴趣度挖掘以及计算作者与主题相关性等功能.模型在挖掘微博主题的同时,可以挖掘出作者的兴趣分布以及作者与特定主题的相关系数,这正是DC-DTM算法中标签传播的依据.

2.2.2 定义及符号说明

在微博文本挖掘的研究中,如果采用标准的LDA模型,如图2所示,通常会由于短文本造成高维系数,导致文本中词之间的相关性不易挖掘,因此,本文首先抽取了一定量的博文进行研究,发现博文通常存在3种产生途径,表现在消息发布方式上就是:原创博文、转发博文以及具有评论性质的博文.原创博文是微博用户发布的原创信息,内容大多是微博用户感兴趣的主题,例如,一个篮球迷可能会发一条这样的博文“科比要退役了,好遗憾!”,而转发博文相对比较特殊,它的内容往往比较简单,在新浪微博中,通过“//@”把转发部分与原创部分相互隔开,而在Twitter中是使用RT标识一篇博文为转发博文,而转发博文的内容往往是被转发博文的链接等.对于对话形式的博文,使用@标识被评论微博的用户,而评论内容相对博文更简单,例如,对上述博文的评论内容可能会是“是呀”.而无论是转发博文还是评论形式的博文,其主题与原博文都具有很强的相关性.如果单独看评论形式的博文,博文中的词语对博文的主题贡献度很小,并且评论形式的博文内容较少,有些只是表明博主的态度等,无法单独从这些博文中进行主题挖掘.而如果将转发或评论形式的博文与原博文连接到一起,就能对转发博文进行主题挖掘,例如上述的评论内容与原博文联系在一起,通过主题挖掘可以发现该评论是关于科比退役这件事情的评论.

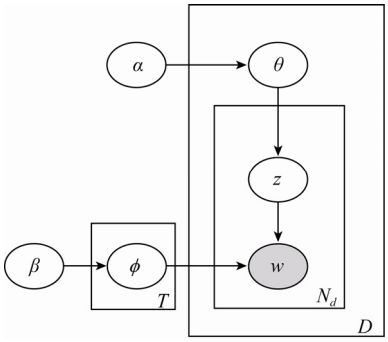

Fig.2 LDA model

图2 LDA模型

上述分析是从博文的内容进行分析,从信息传播层面上,微博中信息的传播主要是通过博主之间相互转发或评论等途径进行.某一博主之所以转发或评论其他博主的博文,是因为对博文的主题或观点感兴趣(这里所指的感兴趣,包括反对或支持等态度),这种对其他微博转发或评论的行为也是博主兴趣点的体现.

根据以上分析,给出如下定义.

定义1(博文). 博主发表的博文,在此包括评论博文与转发博文.如果是评论博文,则将评论博文与被评论博文内容整合在一起,如果是转发博文,并且转发博文的表现形式是链接或其他非原博文形式,则将转发内容与被转发的博文结合在一起.

定义2(用户图). 用户图$GT=(U,E)$.其中,U为博文集合中所有作者的集合,形成用户图GT的节点;E为博文集合中作者之间形成的共现关系,形成用户图GT的边.

此处的用户图即为RT模型得到的图模型,即如果用户u_x的博文评论(或转发)了用户u_y的博文,那么在用户图中,节点u_x和节点u_y将有一条有向边E_{xy}相连,从用户u_x指向用户u_y.

定义3(关注关系). 如果微博b_1评论或转发了微博b_2,那么称b_1关注b_2,同时,微博b_1的作者a_1也关注微博b_2的作者a_2.

定义4. 用户a的被关注概率,是指在某个特定的主题T中,在所有与主题T相关的被关注用户中,用户a占所有被关注用户的百分比.

例如,与主题T相关的关注微博共1万条,由于一个用户可能存在多个关于主题T的博文被关注,假设用户a有100条与主题T相关的博文被关注,那么用户a被关注的概率为100/10000,也就是说,用户在主题T中被关注的概率为0.01.

DTM模型所使用的符号见表1.

Table 1 Symbol description of DTM

表 1 DTM 符号说明

符号	说明
B,B_s,B_r	所有博文集合,原创博文,转发博文
U	所有作者集合
K	主题的数目
V	词表中词的数目
A	作者的数目
N_b	博文 b 中包含的词的数目
W_{bn}	第 b 个博文中第 n 个词
θ_r	转发博文与主题之间的概率分布
θ_s	原创博文与主题之间的概率分布
ϕ_r	转发博文中单词与主题的多项式分布参数
ϕ_s	原创博文中单词与主题的多项式分布参数
φ_r	转发博文中作者与主题的多项式分布参数
φ_s	原创博文中作者与主题的多项式分布参数
$\alpha_r,\alpha_s,\beta_r,\beta_s,\gamma_r,\gamma_s$	模型的先验参数

2.2.3 DTM 模型

DTM 模型是一种基于主题模型的概率生成模型,模型图如图 3 所示.该模型在进行文档主题建模的同时,也进行用户兴趣建模.同时,根据微博所对应的被关注用户概率分布生成微博的被关注者,因此,模型也能对特定主题中的用户被关注情况进行建模.用户的兴趣建模成为计算用户兴趣相似度的基础,用户被关注情况建模成为本文后续的标签传播算法中,标签传播的依据.模型的基本思想是:将每个作者的所有发布、转发或评论的博文都作为文档,每个文档可以表示为词——文档矩阵,在词汇层次中,每个文档都可以表示成一系列主题的混合分布,记为 $P(z)$;同时,每个主题是词汇表中所有单词上的概率分布,记为 $P(w|z)$;在作者层次中,每个文档可以看做是词——作者矩阵,同时,由于在本模型中,作者也具有语义信息,某个作者代表了其所感兴趣的主题,因此每个主题又是作者集合中所有作者上的概率分布,记为 $P(u|z)$.

DTM 模型将所有微博作为语料库,计算用户的兴趣分布;将转发微博提取出来,计算某一主题下用户的影响力.

DTM 模型生成文本方式可以用图 3 所示的贝叶斯网络图来表示.首先,某个文档以一定的概率选择某些主题,这些主题又以一定的概率选择词语和用户.主题是基于词语 w 的概率分布,也是基于用户 a 的概率分布.最初,DTM 从参数为 β的狄利克雷(Dirichlet)分布中抽取主题与单词的关系ϕ;从参数为γ的 Dirichlet 分布中抽取主题与用户之间的关系φ.DTM 生成一条微博时,首先判断微博类型,如果是关注微博,则从参数为α_r的 Dirichlet 分布中抽样出该文档与各主题之间的关系θ_r;否则,从参数为α_s的 Dirichlet 分布中抽样出该文档与各主题之间的关系θ_s.然后,判断文本之间的关系.如果微博为关注微博,则从参数为θ_{dr}的多项式分布中抽样出当前单词所属的主题 z_{dr},从参数为ϕ_{zdm} 的多项式分布中抽取出具体单词.如果是原创微博,则从参数为θ_{ds} 的多项式分布中抽样出当前单词所属的主题 z_{ds},最后从参数为ϕ_{zdsn} 多项式分布中抽取具体单词.

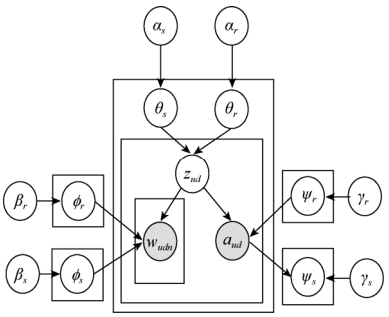

Fig.3 DTM model

图 3 DTM 模型

一条微博中,所有单词与所属主题的联合概率分布如式(2)所示.

$$P(w,z\mid\alpha,\beta)=P(w\mid z,\beta)P(z\mid\alpha) \tag{2}$$

同理,一个文档中所有用户与其所属主题的联合概率分布如式(3)所示.

$$P(a,z\mid\alpha,\gamma\ )=P(a\mid z,\gamma)P(z\mid\alpha) \tag{3}$$

其中,变量 w,z,α,β,γ根据博客的类型进行选择,如果是原创微博,则选择与原创微博相关的参数;如果是转发微博,

则选择转发微博相关变量.上述过程最终得到用户兴趣的主题分布矩阵.

为了对某一主题下用户的影响力建模,该模型将所有转发微博提取出来,对于与主题 T 相关的转发微博 D_r,用户 U 生成主题多项式分布 θ_r,θ_r 服从以 α 为参数的狄利克雷分布.对于用户 u 的每条关注微博 $d_r \in D_r$,根据 θ_r 为这条关注微博 d_r 选取主题 Z_{udr},根据主题 Z_{udr} 所对应的词分布多项式分布参数 ϕ_k 可生成微博 d_u 中的词语 w_{dun},根据主题 Z_{ud} 所对应的被关注用户多项式分布参数 φ_r 可生成微博 d_r 的被转发用户 a_{udr},因此 DTM 模型可同时建模微博主题词的概率分布以及主题的用户被转发概率分布.

假设整个转发微博集合中共有|K|个主题,DTM 模型生成过程如下.

Step 1. 对于每个主题 $k \in K$,选取 $\Phi_r \sim Dir(\beta_r)$,$\Phi_s \sim Dir(\beta_s)$,$\varphi_r \sim Dir(\gamma_r)$,$\varphi_s \sim Dir(\gamma_s)$.

Step 2. 将用户集合按照所发博文类型分为发布博文的用户 U_s 和关注他人博文的用户 U_r.

Step 3. 对于原创博文用户 u,用户集合 U_s 中的每个用户 $u \in U_s$ 选取 $\theta_s \sim Dir(\alpha_s)$.

对于被关注的用户 a_{uds},选取 $a_{uds} \sim Multinomial(\varphi_s)$.

对于用户 u 的每条原创博文 $d \in D$,选择主题 $Z_{ds} \sim Multinomial(\theta_s)$,其中 Z_{ds} 对应主题集合 K 中的主题 k.

对于微博 d 中的第 N_{ud} 个单词 w_{ud},选取 $w_{ud} \sim Multinomial(\Phi_s)$.

Step 4. 对于关注的博文用户 u_r,用户集合 U_r 中的每个用户 u_r,选取 $\theta_r \sim Dir(\alpha_r)$.

对于被关注的用户 a_{ud},选取 $a_{ud} \sim Multinomial(\varphi_r)$.

对于用户 u_r 的每条关注博文 $d_r \in D_r$,选取主题 $Z_{dr} \sim Multinomial(\theta_r)$,主题 Z_{dr} 对应主题集合 K 中的主题 k_r.

对于微博 d_r 中的第 N_{udr} 个单词 w_{udr},选取 $w_{udr} \sim Multinomial(\Phi_r)$.

2.2.4 模型参数估计

DTM 模型的推导采用 Gibbs Sampling 的采样算法,计算模型参数 α_s,α_r,β_s,β_r,γ_r 和 γ_s.词与主题以及作者与主题的联合分布公式如式(4)、式(5)所示.

$$P(\theta, w, z|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \tag{4}$$

$$P(\theta, a, z|\alpha, \gamma) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(a_n|z_n, \gamma) \tag{5}$$

其中,α,β,γ,θ 根据微博是原创还是关注微博,取相应的参数.采用 Gibbs Sapling 采样算法来近似推导 DTM 的参数,得到式(6)、式(7).

$$P(z_i = j|w, z_{-i}, \alpha, \beta) = \frac{P(z, w|\alpha, \beta)}{P(z_{-i}, w|\alpha, \beta)} \propto \frac{n_{j,w} + \beta - 1}{n_{j,\cdot} + V\beta - 1} \times \frac{n_{d,j} + \alpha - 1}{n_{d,\cdot} + T\alpha - 1} \tag{6}$$

同理,

$$P(z_i = j|a_i, z_{-i}, \alpha, \gamma) \propto \frac{f_{-i,j}^{(a_i)} + \gamma}{f_{-i,j}^{(\cdot)} + |a|\gamma} \frac{f_{-i,j}^{(d_i)} + \alpha}{f_{-i,\cdot}^{(d_i)} + T\alpha} \tag{7}$$

其中,$n_{j,w}$ 表示单词 w 属于主题 j 的个数,$n_{j,\bullet}$ 表示属于主题 j 的单词总数;$n_{d,j}$ 表示文档 d 属于主题 j 的词数,$n_{d,\bullet}$ 表示文档 d 所有被分配主题的词数;$f_{-i,j}^{(a_i)}$ 表示除当前样本外,作者 a_i 对主题 j 感兴趣的次数,$f_{-i,j}^{(\cdot)}$ 表示除当前样本外,对 j 感兴趣的作者总数;$f_{-i,j}^{(d_i)}$ 文档 d_i 中关注主题 j 的作者个数,$f_{-i,\bullet}^{(d_i)}$ 表示文档 d_i 中的作者总数;|a|表示作者总数目,T 表示话题总数.

模型最后将某一作者的转发微博的主题分布与原创微博的主题分布进行加和归一,即为作者的主题概率分布.

2.3 节点的微博主题相似度及作者影响力计算

通过第 2.2 节得到的主题——用户分布矩阵,u_1 与 u_2 之间的相似度可以通过计算其主题分布 p_{u1} 和 p_{u2} 相似度进行计算,分布的相似度可以有许多计算方式.KL 散度[30]是衡量不同分布 p 和 q 之间相异度的主要指标,如式(8)所示.

$$KL(p,q)=\sum_{j=1}^{T} p_j \log_2 \frac{p_j}{q_j} \tag{8}$$

为了体现作者之间的相似度越高,作者之间的连边权重越大,本文使用 $D(p,q)$计算作者之间的相似度,如式(9)所示.

$$D(p,q)=1-\sum_{j=1}^{T} p_j \log_2 \frac{p_j}{q_j} \tag{9}$$

其中,如果 p_j 与 q_j 相等,则 $KL(p,q)=0$,$D(p,q)=1$.

在 DTM 模型中,得到每个特定主题 T 下,作者被关注的概率分布 $\varphi_r.\varphi_{ra}$ 表示 a 在主题 T 中的被关注概率,即为用户 a 在主题 T 中的影响力.将用户在某一主题中的影响力按照从大到小排序,此顺序即为之后的标签传播算法 WLPA 中标签传播的顺序.

2.4 基于标签传播的社区发现算法 WLPA

针对 RT 模型,在利用 DTM 模型的进行微博主题挖掘并计算出博主之间的兴趣相似度之后,本文提出基于有向加权网络的标签传播算法 WLPA 进行挖掘.挖掘过程分为 4 个阶段.

(1) 初始化阶段,网络中所有节点的最大标签数即为 DTM 模型中挖掘出的主题个数 k,也就是每个节点所隶属的最大社区个数.用 P_i 表示节点 i 的标签矩阵,p_{ij} 表示节点 i 属于第 j 个社区的概率,即 DTM 中用户——主题概率分布,即用户对不同主题感兴趣的程度.

(2) 在标签传播阶段,标签更新过程中,按照某一主题中用户的影响力排序进行更新,标签传播过程是一个迭代的过程,假设在第 s 次迭代过程中,对于网络中的某一节点,考虑所有指向节点的边,假设节点 i 有 h 个邻接节点,每条链路的权重为 $w_t,t\in\{1,2,\ldots,h\}$,该节点的标签更新矩阵计算如式(10)所示.然后对 p_{ij} 归一化处理,最后删除 $p_{ij}<1/k$ 的标签.每次迭代,根据节点的更新顺序重新计算各节点的标签分布.

$$p_{ij}=\sum_{t=1}^{h}\left(\frac{w_t}{\sum_{t=1}^{h} w_t}\times p_{ij}\right) \tag{10}$$

(3) 经过迭代计算后,网络中各个节点的标签不再变化,达到稳定状态,迭代结束.

(4) 最后,根据每个节点的标签分布向量 C_i 将具有相同标签的节点划为同一社区.在所得到的社区中删除重复节点个数多的社区,最后得到重叠社区结构,其中某些节点拥有多个标签,这些节点即为重叠节点.

算法 1. WLPA 算法描述.

输入:加权有向网络 $G(V,E)$,参数 k,网络总节点数 n.

输出:重叠社区 C_{ov}.

```
For  v∈V do
     Initialize P_v
End for
t=0   //迭代次数
while (!converged()) do
     sort(v)   //对节点按照影响力排序
     t=t+1
     for  v∈V  do
          for  j∈k  do
               calculate p_ij
               if max(p_ij)>1/k
                    p_ij =max(p_ij)
               else if p_ij<1/k
```

p_{ij}=0
end if
normalize(p_{ij})
end for
end for
end while
C=*postProcessing*() //后续处理
Return C

其中,函数 *converged*()用于判断迭代是否收敛.本算法中,如果具有相同标签的节点个数不再变化,则认为算法收敛,*converged*()返回值为真,算法迭代结束.函数 *postProcessing*()用于进行算法的后续处理,即将具有相同标签的节点归为一个社区.

2.5 算法复杂度分析

关于本算法的复杂度分析:假设网络有|*V*|个节点、|*E*|条边、节点归属的最大社区数 *C*、主题个数 *K* 以及文档中词的总数 *N*.

(1) DTM 模型的算法复杂度与 LDA 模型的算法复杂度都与主题个数以及文档中词的总数有关,单次迭代的复杂度为 $O(K\cdot N)$.

(2) 标签传播算法复杂度:第 1 步初始化标签分布矩阵复杂度为 $O(C\cdot|V|)$;第 2 步获取节点的影响力以及节点之间的相似程度的复杂度:为了降低算法复杂度,本文只计算单向的相似度,如在主题 *T* 中,作者 *a* 的关注度大于作者 *b* 的关注度,则只计算 *KL*(*a*,*b*)而不计算 *KL*(*b*,*a*),因此,此步骤的复杂度为 $O(C^2\cdot|V|)$;第 3 步,整个传播过程复杂度 $O\left(C\cdot|E|\cdot\log\left(\frac{C\cdot|V|}{|E|}\right)\right)$;第 4 步,划分社区需要 $O(C\cdot(|E|+|V|))$.由于与|*V*|相比,*C* 取值很小,又由于微博网络为大型稀疏网络,|*E*|的取值相对于|*V*|也很小,因此,WLPA 算法的总时间复杂度为 $O(|V|\cdot\log|V|)$.

DC-DTM 算法复杂度为 $O(K\cdot N+|V|\cdot\log|V|)$,算法复杂度具有线性特征,在真实网络的实验中能够快速收敛.

3 实验及结果分析

3.1 实验准备

3.1.1 数据集

本文采用的数据集来源于新浪微博.该数据集收集了 2013 年 1 月~2014 年 2 月之间的 948 648 条微博内容.首先对该内容进行实验准备:将数据集中转发或评论微博与原微博结合在一起,而原微博保持不变.本文选取了 160 748 个联系人的博文作为实验数据.通过数据预处理,选择出 10 个主题社区:电影、美食、IT、母婴、文艺漫画、电视台主持人、经济、休闲、健身、官方微博作为测试集,其他博文作为模型的训练集.

3.1.2 数据预处理

由于数据集本身包含的原始数据没有用户关系信息,并且原始的博文包含所有词汇,不适合直接使用 DTM 模型进行分析.在分析之前,从具有@关系或//@关系的博文中抽取用户之间的关系形成用户图.对中文语料进行分词,利用中国科学院的分词工具对微博数据进行去停用词、清洗数据以及中文分词等操作.

3.1.3 参数设置

困惑度(perplexity)[1]是一种信息理论的测量方法,是主题模型评价的常用方法.某一个概率模型的 perplexity 值定义为基于该概率的熵的能量,表示预测数据的不确定性.模型的 perplexity 值越小,表示主题建模的效果越好,模型的推广性越高.

Blei 的论文里列出了 LDA 语言模型的 perplexity 的计算公式,如式(11)所示.

$$perplexity(D)=\exp\left\{-\frac{\sum_m \ln p(d_m)}{\sum_m N_m}\right\} \tag{11}$$

其中,D 为测试集,d_m 为每个测试文档,N_m 为第 m 篇文档的大小(即单词个数).

由于模型的性能与参数有关,首先对模型参数进行实验,将 6 个超参取值为[0.1,0.9]之间进行组合,分别观察不同情况下,模型的困惑度.经过多次实验,结果表明,当$\alpha_r=\alpha_s$=0.5 时模型困惑度较小,在此情况下进一步实验,实验数据见表 2.实验结果表明,当$\alpha_r=\alpha_s$=0.5,$\beta_r=\beta_s$=0.3,$\gamma_r=\gamma_s$=0.2 时,模型困惑度最小.接着在上述参数设定的情况下,通过改变主题个数对模型进行实验,实验结果如图 4 所示,当 T=8 时,模型的困惑度达到最低值,也就是说,主题划分最合理.通过对实验数据的分析发现,将健身主题与休闲主题看作同一个主题分类,将文艺漫画与电影划分为一类,因此,原来的 10 类微博数据被分为 8 类.通过迭代次数的实验,如图 5 所示,当迭代次数为 800 时,模型收敛,模型的困惑度趋于常量,因此,在后续实验中,模型中的采样迭代次数为 800 次.

Table 2 Perplexity of DTM model with α=0.5

表 2 参数α=0.5 时 DTM 模型的困惑度

α	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
β	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
γ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Perplexity	68.67	69.40	52.499	68.896	73.369	73.762	75.542	74.212	93.619

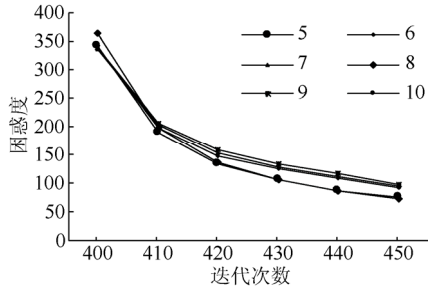

Fig.4 Perplexity of DTM model

图 4 DTM 模型的困惑度

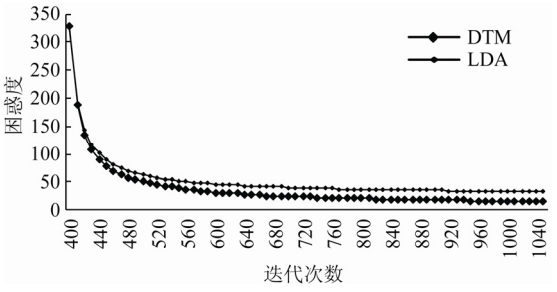

Fig.5 Perplexity of two different model with iterations

图 5 不同迭代次数两个模型的困惑度

3.2 模型分析

3.2.1 模型的困惑度

在上述参数情况下,对比 DTM,LDA 以及 MB-LDA(microblog-latent Dirichlet allocation)模型[24]的内容困惑度发现,在相同参数设置的情况下,DTM 所挖掘出的主题模型的内容困惑度较低,如图 6 所示.

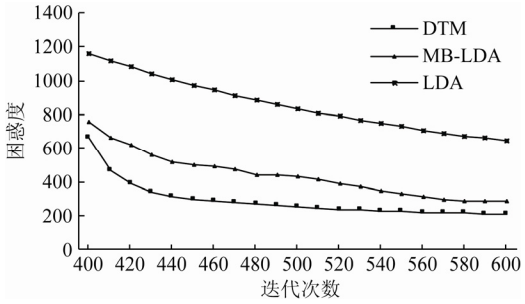

Fig.6 Perplexity of different models with the same parameter

图 6 在相同参数设置情况下模型的困惑度

3.2.2 发现主题的质量

DTM 共挖掘出了 8 个主题,表 3 中展示了 8 个主题.根据各个主题对应的关键词值,分别选择其中 6 个最主要的,从这 6 个词中可以发现,主题 1 与休闲相关,主题 2 与母婴相关,主题 3 与 IT 相关,主题 4 与娱乐相关,主题 5 与经济相关,主题 6 与美食相关,主题 7 与官方微博相关,主题 8 是电视台主持人的微博.DTM 挖掘出的主题基本与我们预先了解到的语料库的主题类似.

Table 3 Top 6 words of 8 different topics

表 3 8 个主题中前 6 个词

TOPIC	主题词
休闲	单反、尼桑、比赛、直播、旅游、出国
母婴	奶粉、宝宝、哭、纸尿裤、海淘、妈妈
IT	苹果、加班、程序猿、人类、小米、三星
娱乐	明星、香港、电影、韩剧、偷拍、公司
经济	创业、股票、国家、政策、经济、投资
美食	美味、制作、吃货、菜、美食、吃
官方微博	日报、晚报、粉丝、官方、登录、微博
电视台主持人	卫视、节目、爱心、关注、传媒、直播

3.2.3 作者兴趣相似度度量

本文中作者兴趣相似度计算方法如式(8)所示.随机选择 10 个作者,其兴趣相似度如图 7 所示.图中颜色越浅,表示作者的相似度越高;对角线上为作者自身的相似度,当值为 1 时,颜色最浅.从图 7 中可以看出,1 号作者与 3 号、5 号、8 号、9 号作者之间的相似度较高,而 10 号作者与 2 号、4 号作者的相似度较低.

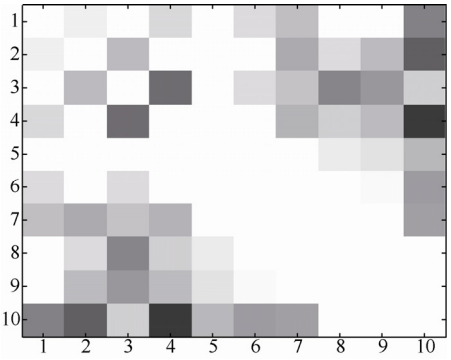

Fig.7 Similarity of different authors' matrix

图 7 作者兴趣相似度矩阵

作者兴趣的相似度可以进一步作为社区发现的基础,通过对比作者之间兴趣的相似度进行作者的社区发现,是一种可调节社区大小,具有层次的重叠社区发现.

3.3 社区发现算法比较

将本文算法与经典的 COPRA 算法进行比较,由于 COPRA 算法解决的是无权网络的社区发现,在此通过将 COPRA 扩充,将其改变为适用于有向加权网络的 WCOPRA 算法.实验选取模型困惑度最小时的社区数目,本数据中,该社区数目为之前实验得到的主题个数,此时模型的拟合程度最好.

3.3.1 加权有向网络社区模块度

传统的社会网络的重叠社区的评价标准,是通过文献[31]所建立的重叠社区的模块度标准进行评价,其公式如式(12)所示.

$$EQ=\frac{1}{R}\sum_{t=1}^{|C|}\sum_{i\in C_t,j\in C_t}\frac{1}{O_iO_j}\left(A_{i,j}-\frac{D_iD_j}{R}\right) \tag{12}$$

其中,D_i为节点d_i的度数,R为网络节点的总度数,A为网络邻居矩阵,O_i为节点d_i所隶属的社区个数.微博网络的社区需要以用户关注关系和用户博文的语义信息作为基础,考虑用户关注关系的有向性,其评价标准 WQ 定义如式(13)所示.

$$WQ=\frac{1}{R}\sum_{t=1}^{|C|}\sum_{i\in C_t,j\in C_t}\frac{S(b_i,b_j)}{O_iO_j}\left(A_{i,j}-\frac{D_{io}D_{jo}}{R}\right) \tag{13}$$

其中,$S(b_i,b_j)$表示用户 i 与用户 j 的兴趣相似度.用户 i 与用户 j 之间的兴趣相似度为两个用户的博文的相似度,由于用户博文随时间增长,并且用户的兴趣也会受到一些实事时间的影响,因此将用户最近一年的博文的相似度作为用户的相似度度量.D_{io}为节点 i 的出度.

首先通过对比算法的模块度进行实验,表 4 为实验结果.结果表明,本文所提出的算法的模块度 WQ 值相对较小.从实验结果可以发现,在使用 EQ 作为评价指标时,由于只考虑微博网络的拓扑结构,并未考虑节点之间的关系,以此为标准作为评价,本文所提出的 DC-DTM 模型表现并不优越,但是 COPRA 与 WCOPRA 两种算法的 EQ 值相似,因为在此 WCOPRA 退化为 COPRA,但当模型的评价指标融入社区节点之间的相似性以及节点之间的关注关系后,DC-DTM 算法比传统的 COPRA 算法以及 WCOPRA 算法都显示出较优的性能.

Table 4 WQ of different algorithms

表 4 不同算法的 WQ

	算法		
	COPRA	WCOPRA	DC-DTM
EQ	0.308	0.309	0.287
WQ	–	0.319	0.382
WDD	–	0.317	0.324

3.3.2 划分紧密度指标

文献[32]提出了划分紧密度(partition density)D 指标,能够有效地衡量重叠社区的划分效果.该指标用来衡量无权无向图的划分.

假设一个无向无权网络含有 M 条边,该网络划分为$\{C_1,C_2,\ldots,C_t\}$个社区,则第 i 个社区的紧密度 D_i 如式(14)所示.

$$D_i=\frac{m_i-(n_i-1)}{n_i(n_i-2)/2-(n_i-1)} \tag{14}$$

其中,m_i,n_i 分别表示第 i 个社区的边数和节点数,当 n_i=2 时,D_i 的值为 0.划分紧密度 D 为所有社区划分紧密度的加权平均,每个社区的权值为社区内的边数与整个网络边数的商,整个网络的划分紧密度如式(15)所示.

$$D=\frac{2}{M}\sum_i m_i\frac{m_i-(n_i-1)}{n_i(n_i-2)/2-(n_i-1)} \tag{15}$$

为了对本文的有向加权网络进行评估,将上述划分紧密度进行改进,得到有向加权网络的划分紧密度指标 WDD,如式(16)所示.

$$WDD=\frac{2}{W}\sum_{i=1}^{t}w_i\frac{w_i-(n_i-1)}{n_i(n_i-2)/2-(n_i-1)} \tag{16}$$

其中,w_i 为社区中所有节点输出边的权重之和,n_i 为属于第 i 个社区的节点数.由于 COPRA 未考虑边权重,在此不做比较.从表 4 可以看出,DC-DTM 算法的划分密度较高.

3.3.3 作者兴趣相似度比较

图 8 为对同一社区内不同作者兴趣相似性的度量.随机选择其中 1 个社区内的用户的兴趣相似度矩阵图,颜色越浅,说明作者兴趣度越相似.从图 8 中可以看出,同一社区内作者兴趣基本相似.图 9 为随机选择不同社区内的作者之间的相似度矩阵图,图中颜色越深,说明作者的兴趣相似度越低.从图 9 中可以看出,作者兴趣基本不同.但是,9 号作者与 2 号、3 号作者的相似度较高,说明他们可能是重叠节点.从图中可以发现,同一社区的作者兴趣相似度较高.不同社区作者的兴趣相似度相对较低.

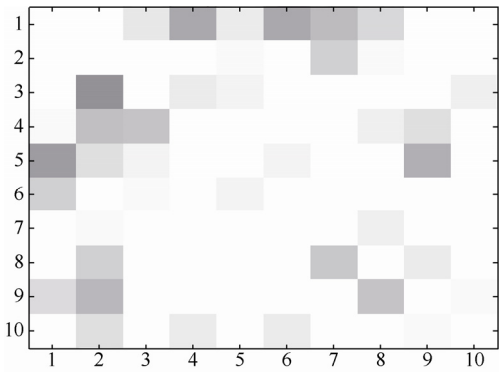

Fig.8 Similarity of different authors in the same community

图 8 同一社区内不同作者的兴趣相似度

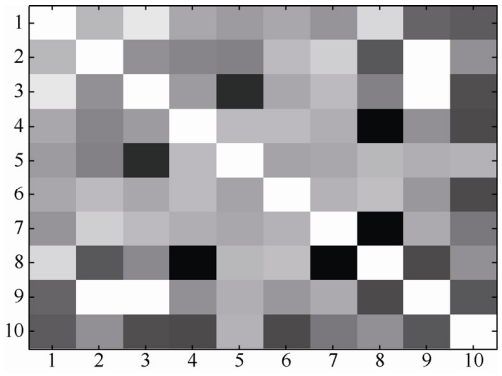

Fig.9 Similarity of different authors in different communities

图 9 不同社区间作者的兴趣相似度

4 结束语

本文针对微博网络的社区划分问题提出了 DC-DTM 算法.在标准的 LDA 模型中,主题由文本本身确定,不适合短文本分析,文中所提出的 DTM 模型引入了作者的主题分布和对转发微博的处理,能够弥补 LDA 模型带来的缺陷.在 DTM 模型中,一篇微博如果是原创微博,则其主题由作者的主题分布中抽取;如果是关注微博,则由原创与关注部分的主题来确定.实验分析验证了本文所提出的 DTM 模型更适合挖掘微博的主题,所提出的 DC-DTM 算法所划分出的社区的模块度较高,社区内作者的兴趣相似度较高,社区之间的作者相似度较低.另外,DC-DTM 算法可为下一步深入研究微博网络的情感分析,大规模社会网络的推荐等研究领域提供帮助.

References:

[1] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003,3:993−1022.

[2] Steyvers M, Smyth P, Rosen-Zvi M, Griffiths T. Probabilistic author-topic models for information discovery. In: Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2004. 306−315. [doi: 10.1145/1014052.1014087]

[3] Salton G, McGill M. Introduction to Modern Information Retrieval. 3rd ed., New York: ACM, 1999. [doi: 10.3724/SP.J.1001.2009.00054]

[4] Yang B, Liu DY, Liu J, Jin D, Ma HB. Complex network clustering algorithms. Ruan Jian Xue Bao/Journal of Software, 2009, 20(1):54−66 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3464.htm [doi: 10.3724/SP.J.1001.2009.00054]

[5] Lin YF, Wang TY, Tang R, Zhou YW, Huang HK. An effective model and algorithm for community detection in social networks. Journal of Computer Research and Development, 2012,49(2):337−345 (in Chinese with English abstract).

[6] Yan B, Gregory S. Detecting community structure in networks using edge prediction methods. Journal of Statistical Mechanics: Theory and Experiment, 2012,2012(9):No.P09008. [doi: 10.1088/1742-5468/2012/09/P09008]

[7] Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005,435(7043):814−818. [doi: 10.1038/nature03607]

[8] Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S. Finding statistically significant communities in networks. Plos One, 2011, 6(4):336−338. [doi: 10.1371/journal.pone.0018961]

[9] Baumes J, Goldberg M, Magdon-Ismail M. Efficient identification of overlapping communities. Lecture Notes in Computer Science, 2005,3495:27−36. [doi: 10.1007/11427995_3]

[10] Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2007,76(3):No.036106. [doi: 10.1103/PhysRevE.76.036106]

[11] Gregory S. Finding overlapping communities in networks by label propagation. New Journal of Physics, 2010,12(10):No.103018.

[doi: 10.1088/1367-2630/12/10/103018]

[12] Liu SC, Zhu FX, Gan L. A label-propagation—Probability-Based algorithm for overlapping community detection. Chinese Journal of Computers, 2016,39(4):717−729 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2016.00717]

[13] Blei D, Ng A, Jourdan M. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003(3):993−1022.

[14] Minka T, Lafferty J. Expectation-Propagation for the generative aspect model. In: Proc. of the 18th Conf. on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, Inc., 2002. 352−359.

[15] Steyvers M, Griffiths T. Probablistic topic model. In: Landauer T, McNamara D, Dennis S, Kintsch W, eds. Latent Semantic Analysis: A Road to Meaning. Springer-Verlag, 2007.

[16] Yao QZ, Song ZL, Peng C. Research on text categorization based on LDA. Computer Engineering and Applications, 2011,47(13): 150−153.

[17] Duan L, Zhu XY. Microblog community detection method based on community spatio-temporal topic model. Journal of University of Electronic Science and Technology of China, 2014,43(3):465−468 (in Chinese with English abstract).

[18] Yang J, Xin Y, Xie ZQ. Semantics social network community detection algorithm based on topic comprehensive factor analysis. Journal of Computer Research and Development, 2014,51(3):559−569 (in Chinese with English abstract).

[19] Xin Y, Yang J, Xie ZQ. An overlapping semantic community structure detecting algorithm by label propagation. Acta Automatica Sinica, 2014,40(10):2262−2275 (in Chinese with English abstract).

[20] Xin Y, Yang J, Xie ZQ. An overlapping community structure detectiong algorithm in semantic social network based on block field. Acta Automatica Sinica, 2015,41(2):362−375 (in Chinese with English abstract).

[21] Xin Y, Yang J, Xie ZQ. A semantic overlapping community detecting algorithm in social networks based on random walk. Journal of Computer Research and Develapment. 2015,52(2):499−511 (in Chinese with English abstract).

[22] Xin Y, Yang J, Xie ZQ. Link-Block method for the semantic overlapping community detection. Ruan Jian Xue Bao/Journal of Software, 2016,27(2):363−380 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4810.htm [doi: 10.13328/j.cnki.jos.004810]

[23] Zhou XP, Liang X, Zhang HY. User community detection on micro-blog using R-C model. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2808−2823 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4720.htm [doi: 10.13328/j.cnki.jos.004720]

[24] Zhang CY, Sun JL, Ding YQ. Topic mining for microblog based on MB-LDA model. Journal of Computer Research and Development, 2011,48(10):1795−1802 (in Chinese with English abstract).

[25] Zhang ZF, Li QD, Zeng D, Gao H. User community discovery from multi-relational networks. Decision Support Systems, 2013, 54(2):870−879. [doi: 10.1016/j.dss.2012.09.012]

[26] Zhang LM, Huang WJ, Chen W, Wang TJ, Lei K. EMTM: A method for experts mining in micro-blog with topic-level. Journal of Computer Research and Development, 2015,52(11):2517−2526 (in Chinese with English abstract).

[27] Hu Y, Wang CJ, Wu J, Xie JY, Li H. Overlapping community discovery and global representation on MicroBlog network. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2824−2836 (in Chinese with English abstract). http:// www.jos.org.cn/1000-9825/4721.htm [doi: 10.13328/j.cnki.jos.004721]

[28] Chai BF, Jia CY, Yu J. Approaches of structure exploratory based on probabilistic models in massive networks. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2753−2766 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4722.htm [doi: 10.13328/j.cnki.jos.004722]

[29] Liu SP, Yin J, Ouyang J, Huang Y, Yang XY. Topic mining from microblogs based on MB-HDP model. Chinese Journal of Computers, 2015,38(7):1408−1419 (in Chinese with English abstract).

[30] Leibler RA, Kullback S. On information and sufficiency. Annals of Mathematical Statistics, 1951,22(1):79−86. [doi: 10.1214/aoms/1177729694]

[31] Shen HW, Cheng XQ, Cai K, Hu MB. Detect overlapping and hierarchical community structure in networks. Physica A: Statistical Mechanics and its Applications, 2009,388(8):1706−1712. [doi: 10.1016/j.physa.2008.12.021]

[32] Ahn YY, Bagrow JP, Lehmann S. Link communities reveal multiscale complexity in networks. Nature, 2010,466(7307):761−764. [doi: 10.1038/nature09182]

附中文参考文献:

[4] 杨博,刘大有,金弟,马海宾.复杂网络聚类方法.软件学报,2009,20(1):54−66 http://www.jos.org.cn/1000-9825/3464.htm [doi: 10.3724/SP.J.1001.2009.03464]

[5] 林有芳,王天宇,唐锐,周元炜,黄厚宽.一种有效的社会网络社区发现模型和算法.计算机研究与发展,2012,49(2):337−345.

[12] 刘世超,朱福喜,甘琳.基于标签传播概率的重叠社区发现算法.计算机学报,2015,39(4):717−729. [doi: 10.11897/SP.J.1016.2016.00717]

[17] 段炼,朱欣焰.基于社区时空主题模型的微博社区发现方法.电子科技大学学报,2014,43(3):464−469.

[18] 杨静,辛宇,谢志强.基于话题综合因子分析的语义社会网络社区发现算法.计算机研究与发展,2014,51(3):559−569.

[19] 辛宇,杨静,谢志强.基于标签传播的语义重叠社区发现算法.自动化学报,2014,40(10):2262−2275.

[20] 辛宇,杨静,谢志强.一种面向语义重叠社区发现的 Block 场取样算法.自动化学报,2015,41(2):362−375.

[21] 辛宇,杨静,谢志强.基于随机游走的语义重叠社区发现算法.计算机研究与发展,2015,52(2):499−511.

[22] 辛宇,杨静,谢志强.一种面向语义重叠社区发现的 Link-Block 算法.软件学报,2016,52(2):363−380. http://www.jos.org.cn/1000-9825/4810.htm [doi: 10.13328/j.cnki.jos.004810]

[23] 周小平,梁循,张海燕.基于 R-C 模型的微博用户社区发现.软件学报,2014,25(12):2808−2823. http://www.jos.org.cn/1000-9825/4720.htm [doi: 10.13328/j.cnki.jos.004720]

[24] 张晨逸,孙建伶,丁轶群.基于 MB-LDA 模型的微博主题挖掘.计算机研究与发展,2011,48(10):1795−1802.

[26] 张腊梅,黄威靖,陈薇,王腾蛟,雷凯.EMTM:微博中与主题相关的专家挖掘方法.计算机研究与发展,2015,52(11):2517−2526.

[27] 胡云,王崇骏,吴骏,谢俊元,李慧.微博网络上的重叠社群发现与全局表示.软件学报,2014,25(12):2824−2836. http://www.jos.org.cn/1000-9825/4721.htm [doi: 10.13328/j.cnki.jos.004721]

[28] 柴变芳,贾彩燕,于剑.基于概率模型的大规模网络结构发现方法.软件学报,2014,25(12):2753−2766. http://www.jos.org.cn/1000-9825/4722.htm [doi: 10.13328/j.cnki.jos.004722]

[29] 刘少鹏,印鉴,欧阳佳,黄云,杨晓颖.基于 MB-HDP 模型的微博主题挖掘.计算机学报,2015,38(7):1408−1419.

刘冰玉(1978－),女,河北秦皇岛人,博士生,讲师,CCF 专业会员,主要研究领域为数据挖掘,复杂网络.

王翠荣(1963－),女,博士,教授,CCF 高级会员,主要研究领域为数据挖掘,网络虚拟化.

王聪(1981－),男,博士,讲师,CCF 专业会员,主要研究领域为网络虚拟化,云资源分配,大数据.

王军伟(1971－),男,博士,教授,博士生导师,主要研究领域为下一代互联网,网络路由协议及算法.

王兴伟(1968－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为未来互联网,云计算,网络空间安全.

黄敏(1968－),女,博士,教授,博士生导师,主要研究领域为物流与供应链管理,智能调度与优化方法.