软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(2):278−291 [doi: 10.13328/j.cnki.jos.005030]

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

基于背景和内容的微博用户兴趣挖掘

仲兆满[1,2], 管 燕[1], 胡 云[1], 李存华[1]

[1](淮海工学院 计算机工程学院,江苏 连云港 222005)

[2](江苏金鸽网络科技有限公司 软件研发中心,江苏 连云港 222005)

通信作者: 仲兆满, E-mail: zhongzhaoman@163.com

摘 要: 微博用户兴趣挖掘是个性化推荐、社群划分的基础工作.在深入分析微博网络特点的基础上,给出了能够揭示微博网络多模性的描述模型,对面向微博网络的后续研究具有参考价值.根据微博网络的特点,提出了基于背景的用户静态兴趣表示及挖掘方法,以及基于微博的用户动态兴趣表示和挖掘方法.针对微博网络中缺少背景信息、发表微博很少的大量不活跃用户,提出了基于关注的用户兴趣挖掘方法.以新浪微博为例,选取了时尚、企业管理、教育、军事、文化这 5 个领域进行用户兴趣挖掘及相似度计算的实验分析和比较,结果表明,与主流的兴趣挖掘方法相比,该微博用户兴趣的表示和挖掘方法可以有效地改善微博用户兴趣挖掘的效果.

关键词: 微博网络;用户兴趣表示;用户静态兴趣;用户动态兴趣;用户兴趣挖掘;用户兴趣相似度计算

中图法分类号: TP311

中文引用格式: 仲兆满,管燕,胡云,李存华.基于背景和内容的微博用户兴趣挖掘.软件学报,2017,28(2):278–291. http://www.jos.org.cn/1000-9825/5030.htm

英文引用格式: Zhong ZM, Guan Y, Hu Y, Li CH. Mining user interests on microblog based on profile and content. Ruan Jian Xue Bao/Journal of Software, 2017,28(2):278−291 (in Chinese). http://www.jos.org.cn/1000-9825/5030.htm

Mining User Interests on Microblog Based on Profile and Content

ZHONG Zhao-Man[1,2], GUAN Yan[1], HU Yun[1], LI Cun-Hua[1]

[1](School of Computer, Huaihai Institute of Technology, Lianyungang 222005, China)

[2](Software Research and Development Center, Jiangsu Jinge Network Technology Co., Ltd., Lianyungang 222005, China)

Abstract: Mining user interests on microblog is the basis for personalized recommendation and community classification. A descriptive model of microblog network is proposed based on the in-depth analysis over the characteristics of microblog in the work, revealing properties of multi-mode microblog. The representation and mining method of profile-based static user interests and microblog post-based dynamic user interests are proposed respectively according to the characteristics of microblog network. For mining inactive users with little profile and few microblog posts, a method of follower-based interest mining is proposed. In the case study of Sina microblog, users in fashion, business management, education, military and culture are selected for experimental analysis and comparison of interest mining and similarity calculation. Experimental results show that the proposed representation and mining method can effectively improve user interest mining comparing with other state-of-the-art methods.

Key words: microblog network; user interest representation; user static interest; user dynamic interest; user interest mining; user interest similarity calculation

微博网络中的个性化推荐、领域专家的发现、社区划分是当前社会计算的研究热点[1−3],而有效地挖掘出微博用户的兴趣取向,是此类研究的基础工作.已有的与用户兴趣挖掘相关的研究总体上可分为两类:基于背景

* 基金项目: 国家自然科学基金(61403156); 江苏省科技厅产学研前瞻性联合研究基金(BY2015048-02)

Foundation item: National Natural Science Foundation of China (61403156); Prospective Joint Research Foundation of University-Industry Cooperation of Jiangsu (BY2015048-02)

收稿时间: 2015-08-29; 修改时间: 2015-12-02; 采用时间: 2016-01-22

的用户兴趣挖掘和基于内容的用户兴趣挖掘.

(1) 基于背景的用户兴趣挖掘

不同的社交平台用户背景的描述会有所不同,但基本包含以下类别:个人简介、标签、职业、毕业院校、出生地、出生日期、性别等信息.

Wang 等人[4]在研究重叠社区发现时认为,用户的关联性(粉丝或关注)过于自由,重点使用了用户的元数据Metadata(比如标签)提取用户的兴趣.Diaby 等人[5]研究社交网络推荐时,考虑的是用户的背景信息,对不同的社交媒体,选取了不同的背景信息,主要包括工作、教育、简历、兴趣、职位等.进一步地,利用了用户的朋友(friend)信息,但结论是背景相似的朋友才有价值.文献[6−8]在社交推荐系统中也都有朋友信息的利用,主要是朋友的背景信息.文献[9,10]在研究社区发现时认为,使用用户背景、共享图片、视频和标签等信息既简单又有效.Ghosh等人[11]根据 Twitter 中用户对其所关注的对象添加分组描述的信息,通过收集多个用户对同一个用户的分组描述信息,使用出现最多的部分描述词作为用户的描述.在 Liang 等人[12]的工作中,标签被看作是微博用户对其自身专长领域的描述,使用了一个公开的流言数据集,通过人工标注出与每条流言相关的专家用户.结果表明,基于标签的方法效果好于基于微博内容的语言模型方法.Akcora 等人[13]在计算社交网络用户的相似度时综合用户的背景信息和网络结构,在 Facebook 平台统计发现,64%的用户缺少背景信息的描述,提出了从用户朋友已有的数据中自动推理出用户的一些可能的背景信息.邢千里等人[14]也是围绕用户的个人描述进行用户的兴趣提取,但不同的是,他们认为微博的标签在描述用户兴趣方面的利用价值更大,详细地分析了微博中用户添加标签的行为及标签内容分布的特点,实验验证了基于标签的预测方法其效果优于基于微博内容的预测方法.

基于背景挖掘用户兴趣存在的问题:① 用户兴趣是通过多方面反映的,仅仅通过背景难以全面地反映用户兴趣,尤其是一些短期的话题,比如对研究者而言,当发生“院士造假”此类突发事件时,用户可能在短期内会深度关注,而这些话题在背景中难以反映;② 用户背景信息在很多情况下是不完善的,仅基于背景难以完成用户兴趣挖掘的目的,比如文献[14]对新浪微博 263 万用户进行统计,发现 59.4%的用户没有添加标签,7%的用户只是象征性地添加了一个标签,文献[15]对新浪微博 1.4 亿用户进行统计,发现 78.2%的用户没有添加标签,标签数小于 5 的用户占用户总数的 93.8%.

(2) 基于内容的用户兴趣挖掘

用户在各类社交平台上经常会发表、评论或者转发大量的信息,从这些信息中能够挖掘出用户的兴趣取向.

Ma 等人[16]提出从多个数据源挖掘用户兴趣,用户的兴趣用若干关键词表示,数据源是指用户在不同平台上发表的帖子,如 Twitter,Facebook,Linkedin 等.Chen 等人[17]比较了使用用户自己发表的微博构建用户的兴趣词袋和使用用户的粉丝构建用户的兴趣词袋两种方法,发现前者效果更好.Weng 等人[18]将每个用户发表的所有微博合并成一个大的文档,然后使用标准的 LDA 模型在文档中提取用户兴趣.Zhao 等人[19]认为微博比较短小,一条微博中的所有单词仅有 1 个主题生成,即 1 条微博对应 1 个主题.周小平等人[20]在研究微博用户社区发现时,定义了关注关系的兴趣特征为其所关联的两个用户的兴趣特征的共同部分,形成了兴趣和网络结构双内聚的用户社区发现方法,用户的兴趣特征提取来源于用户发表的微博内容.Syeyvers 等人[21]认为主题是多个关键词的概率分布,用户也以某种概率分布对多个主题感兴趣,并提出了 AT(author-topic)模型,用于发现用户、文档、主题和关键词之间的关系.Zhang 等人[22]综合了用户主题模型(AT)和用户关系网络研究用户的社区发现,用户主题模型的构建基于用户在社交网络上发表的内容,并在 Tweets 和 Delicious 上进行了验证.彭泽环等人[23]在研究微博用户推荐时,考虑了用户发表的微博信息,但没有进一步区分内容中的话题、标题的权重.

基于内容挖掘用户兴趣存在的问题:① 社交平台上有很多“冷启动”用户,此类用户可能是新注册的用户,也可能是不活跃用户,仅基于内容难以挖掘到此类用户的兴趣;② 用户发表的信息都是随着时间而动态变化的,有的兴趣是长期的,有的兴趣是短期的,已有的研究方法未能体现社交平台用户兴趣的动态性;③ 在用户发表的微博中,除正文外,可能还包含话题、标题等信息,用户发表微博有原创、转发及评论等不同的方式,已有研究对此分析应用不够.

本文在总结了已有研究工作不足的基础上,提出的微博用户兴趣挖掘方法的创新点包括:① 给出了微博网络的描述模型,该模型深入地揭示了微博的多模特性,对面向微博网络的后续研究具有重要参考价值;② 根据微博网络的特点,提出了基于背景的用户静态兴趣表示及挖掘方法,以及基于微博内容的用户动态兴趣表示和挖掘方法,在微博内容的分析时,考虑了其中的话题、标题的信息价值,同时也区分了用户发表、转发及评论微博的不同情况;③ 针对微博网络大量的用户缺少背景、发表微博很少的不活跃用户,提出了基于关注的兴趣挖掘方法,综合地考虑了关注用户的"等级信息"及用户间的交互强度进行关注的选取.

1 微博网络模型

已有的研究在描述微博网络时,仍然以传统的二部图为主,将微博网络抽象成用户-话题模型,二部图 $G=(V,E)$将节点分为两个互不相交的子集$\{V_1,V_2\}$,并且图中的每条边 $e_i\in E$ 所关联的两个节点分别属于 V_1 和 V_2. TwitterRank 模型[13]将用户发表的所有微博合并成一个文档,在文档中提取主题,构建微博的用户和主题表示模型,文献[24]将该方法称为用户视图.TwitterLDA 模型[19]认为一条微博中的所有单词仅有 1 个主题生成,是基于用户视图的扩展主题模型.虽然有个别文献提到了微博多模网络的概念,比如文献[25],但是其仍然采用了用户-话题的机制,但未能揭示微博网络的微博之间、用户之间、微博与用户之间真实存在的各种复杂关系.

本文根据微博媒体的特点提出了微博网络模型,如图 1 所示.

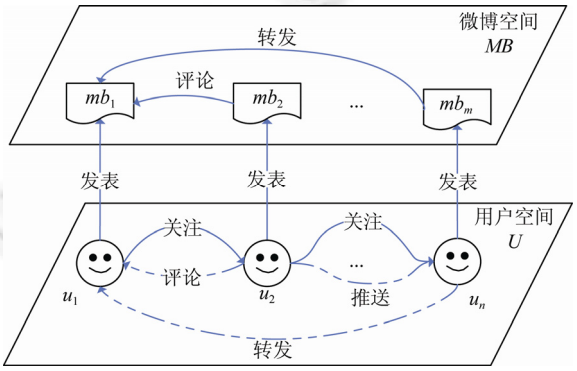


Fig.1 Microblog network model

图 1 微博网络模型

由图 1 可见,微博网络的核心是两个空间:用户空间和微博空间.在这两个空间的基础上,用户与微博之间、用户之间、微博之间形成了各种关系,是一种典型的复杂网络.比如,用户 μ_1 发表了微博 mb_1,用户 μ_n 发表微博 mb_m 时对微博 mb_1 进行了转发,这时微博 mb_m 和 mb_1 是一种转发关系,同时用户 μ_n 和 μ_1 也基于发表的微博构建了转发关系.微博网络模型的相关概念解释如下.

定义 1. 微博网络,描述为一个九元组: $MBN=(U,MB,E_{UMB},E_{MBC},E_{MBF},E_{UU},E_{UForU},E_{UCU},E_{UPU})$, 其中, $U=\{u_1,u_2,...,u_n\}$ 为用户集; $MB=\{mb_1,mb_2,...,mb_m\}$ 为微博集; $E_{UMB}=\{e=(u_i,mb_j)|u_i\in U,mb_j\in MB\}$ 为用户与其所发表微博的关系集; $E_{MBC}=\{mb_i\rightarrow mb_j)\,|\,mb_i,mb_j\in MB,mb_i \text{ comments } mb_j\}$ 为微博之间的评论关系集; $E_{MBF}=\{mb_i\rightarrow mb_j)\,|\,mb_i,mb_j\in MB,mb_i \text{ forwards } mb_j\}$ 为微博之间的转发关系集; $E_{UU}=\{u_i\rightarrow u_j)\,|\,u_i,u_j\in U,u_i \text{ follows } u_j\}$ 为用户通过关注关系而形成的连接关系集; $E_{UForU}=\{(u_i,u_j)\,|\,u_i,u_j\in U,u_i \text{ publishes } mb_i,u_j \text{ publishes } mb_j,mb_i \text{ forwards } mb_j\}$ 为用户通过微博之间的转发关系而形成的用户间的转发关系集; $E_{UCU}=\{(u_i,u_j)\,|\,u_i,u_j\in U,u_i \text{ publishes } mb_i,u_j \text{ publishes } mb_j,mb_i \text{ comments } mb_j\}$ 为用户通过微博之间的评论关系而形成的用户间的评论关系集; $E_{UPU}=\{(u_i,u_j)\,|\,u_i,u_j\in U,u_i \text{ publishes } mb_i,u_i \text{ pushes } mb_i \text{ to } u_j\}$ 为用户发表时推送给其他用户而形成的关系集.

基于定义 1 微博网络的概念,既可以从用户空间的角度出发,也可以从微博空间的角度出发,对微博网络做

很多有价值的分析研究.比如,研究用户之间的交互强度、转发或评论关系、相似性,研究微博中的话题提取、微博之间的转发或评论关系、微博源头的追踪,还可以从用户和微博融合的角度做用户的兴趣挖掘、社区划分、微博热点挖掘等.

定义 2. 微博博文,简称微博,描述为一个三元组:$MBlog_i$=(*body*,*t*,*u*),其中,*body* 为微博主体内容,*t* 为微博发表的时间,*u* 为发表该微博的用户.

定义 3. 微博用户,描述为一个六元组:u_i=(*name*,*profile*,*MB*,*follower*,*fans*),其中,*name* 为微博的用户名,是微博网络中用户的唯一标识符;*profile* 为微博平台上的用户背景,不同微博平台背景有所差异;*MB* 为用户在微博网络上发表的微博集;*follower* 为用户的关注集;*fans* 为用户的粉丝集.

2 基于背景和内容的微博用户兴趣挖掘

2.1 微博用户兴趣表示

已有研究定义的用户兴趣表示模型见定义 4.

定义 4. 用户兴趣[16,19−21],普遍被定义为用户对各个兴趣点的喜好程度,$UI=\{Int_1,Int_2,\ldots,Int_m\}$,每个兴趣点是一个二元组 $Int_i=(topic_i,w_i)$,$topic_i$ 为话题,通常由多个关键词组成;w_i 为用户对 $topic_i$ 的喜好权重.假设用户 u_i 有两个兴趣点 $Int_1=\langle topic_1,w_1\rangle=\langle\{$军事,飞机,性能$\},0.8\rangle\}$,$Int_2=\langle topic_2,w_2\rangle=\langle\{$旅游,户外,爬山$\},05\rangle\}$,由于权重 $w_1=0.8>w_2=0.5$,这意味着这个用户更喜欢 $topic_1$.

在进行微博用户兴趣的描述之前,先分析微博用户兴趣的来源.微博用户兴趣来源于两处:① 用户背景,不同的微博网络用户的背景会有些差异,但简介、标签、职业(行业)类信息都能很好地反映用户的兴趣.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.背景代表了用户的总体偏好,是经过长时间积累形成的,体现为一种静态兴趣、长期兴趣,比如用户的研究领域、爱好特长等;② 用户发表的微博,包括用户发表、评论、转发的各类微博,用户的微博直观地反映了用户产生信息的兴趣偏好.微博体现为用户的动态兴趣,是长期兴趣和短期兴趣的结合,短期兴趣则相对不稳定,会不定期地变化.例如,在世界杯期间,用户可能会对世界杯感兴趣;天津发生大爆炸事件,用户可能在一段时间范围内经常参与讨论.再如,用户围绕自己的研究领域,会经常发表相关的微博,这些内容是用户的长期兴趣.

因此,挖掘微博用户兴趣时,应该既有来源于背景的静态兴趣,又有来源于微博的动态兴趣,只有这样才能合理地描述微博用户的兴趣.

定义 5. 微博用户静态兴趣是指从用户背景中挖掘出的兴趣点,$UI=\{Int_1,Int_2,\ldots,Int_m\}$,每个兴趣点是一个二元组 $Int_i=(kw_i,w_i)$,kw_i 为关键词;w_i 为用户对 kw_i 的喜好权重.假设用户 u_i 有两个兴趣点 $Int_1=(kw_1,w_1)=$(信息检索,0.5),$Int_2=(kw_2,w_2)=$(旅游,0.2),由于权重 $w_1=0.5>w_2=0.2$,这意味着这个用户更喜欢 kw_1.

定义 5 和定义 4 非常类似,都是若干的兴趣点及其权重.但不同的是,由于从一个用户简短的背景中提取兴趣点时,难以进行多个关键字之间的有效聚类,因此定义 5 中的每个兴趣点都只包含 1 个关键词.

定义 6. 微博用户动态兴趣是指从用户微博中挖掘出的随时间变化而变化的兴趣点,$UI=\{Int_1,Int_2,\ldots,Int_m\}$,每个兴趣点为一个三元组 $Int_i=(topic_i,w_i,T)$,其中,$topic_i$ 是由多个关键词组成的话题;w_i 为用户对 $topic_i$ 的喜好权重;$T=\{t_1,t_2,\ldots,t_s\}$,t_i 为用户讨论话题 $topic_i$ 的各个时间点,即话题在不同时间点的分布情况.

定义 6 在表示微博用户的兴趣时引入了时间分布的思想,除了能够体现用户的兴趣点及其权重之外,还能反映用户在不同时间段的兴趣,这种表示模型将有助于深入分析用户的兴趣特征.Zhao 等人[26]在识别突发事件时引入了突发特征词(bursty word)的概念,给出了特征词的起始时间 t_s 和结束时间 t_e,根据 t_s 和 t_e 仅能计算特征词的时间跨度,无法获取特征词的时间持续度.定义 6 给出的微博用户动态兴趣表示模型根据时间粒度的不同(假设时间粒度为“天”),既可以计算用户兴趣的时间跨度($|t_s-t_1|$),又可以计算用户兴趣的时间持续度($|T|$),还可以统计兴趣点关于时间的分布情况.

如果不考虑用户兴趣点的时间分布,仅从兴趣点的角度出发,在计算用户兴趣点的重要度时,短期兴趣很可能会“淹没”长期兴趣.比如,用户在仅 1 天内就话题 $topic_1$ 讨论了 20 次,以后不再谈及;而在一周的每一天都讨论

了话题 $topic_2$,但总数为 15 次,如果不考虑时间持续度,$topic_1$ 的权重明显比 $topic_2$ 大.但从更能体现用户的稳定兴趣而言,长期兴趣更能体现用户的真实兴趣取向,这时 $topic_2$ 的权重应该大于 $topic_1$.

2.2 微博用户兴趣挖掘框架

依据从微博网络中挖掘用户静态兴趣和动态兴趣的需求,本文提出的微博用户兴趣挖掘框架如图 2 所示.

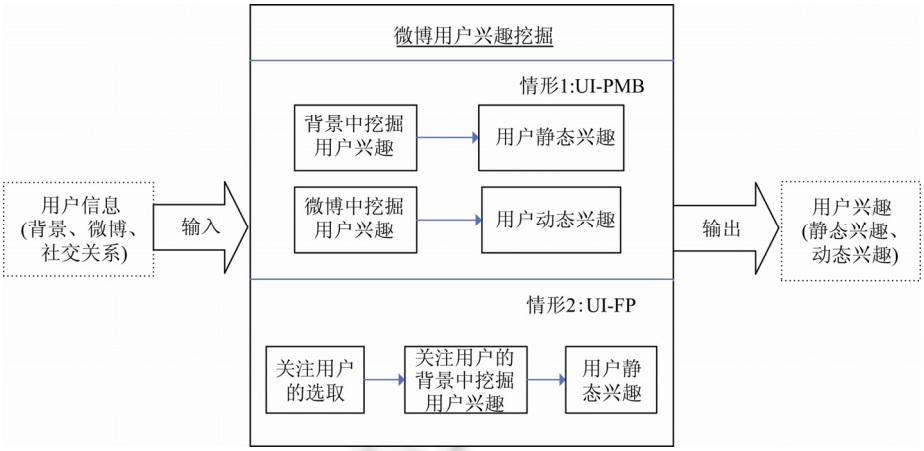


Fig.2 Framework of microblog user interest mining

图 2 微博用户兴趣挖掘框架

图 2 所示的微博用户兴趣挖掘框架由 3 部分组成:① 用户信息,包括用户背景、发表的微博(原创、转发或评论)、社交关系(关注、粉丝及其交互),这些内容使用网络信息采集工具可以方便地获取;② 用户兴趣挖掘,分为两种情形:情形 1 指用户背景或者用户微博信息量较大,能够直接从中挖掘出用户的兴趣,这种情形简记为 UI-PMB;情形 2 指用户背景和用户发表的微博信息量都较小,难以从这些信息中直接挖掘出用户的兴趣,采用从关注中挖掘用户兴趣的替代方法,这种情形简记为 UI-FP;③ 用户兴趣,参考定义 5 和定义 6 的论述,输出用户的静态兴趣和动态兴趣.

本文第 2.3 节和第 2.4 节将详细介绍 UI-PMB 和 UI-FP 两种情形下的用户兴趣挖掘方法.

2.3 UI-PMB情形用户兴趣挖掘

2.3.1 基于背景的用户静态兴趣挖掘

本文以新浪微博的用户背景为例,介绍微博用户背景的兴趣挖掘方法.新浪微博能够体现用户兴趣的背景,包括简介、标签、职位等信息.

用户在新浪微博定义自己的标签时,既可以在微博网络的标签库中选取,也可以人工输入.标签可以理解为有意义的字符串,把实验采集到的用户的标签形成一个库,在分词时,将其作为一个整体识别.用户 u_1 的标签记为 $u_1.tag=\{tag_1,tag_2,...,tag_e\}$.

对用户的简介分词,因为内容较短,不进行词频统计,得到用户 u_1 的简介词集为 $u_1.bi=\{bi_1,bi_2,...,bi_f\}$.采用同样的策略,得到用户 u_1 的职位词集为 $u_1.job=\{job_1,job_2,...,job_g\}$.

统计 $u_1.tag$,$u_1.bi$ 及 $u_1.job$ 中各个词出现的次数,作为词的权重,权重参考最大值进行归一化,按照权重降序排列,根据需求选取 m 个词作为用户的静态兴趣,记为 $UPI=\{(kw_1,w_1),(kw_2,w_2),...,(kw_m,w_m)\}$.

2.3.2 基于微博内容的用户动态兴趣挖掘

从微博中提取用户的兴趣(话题)是近期的热点研究问题.思路之一是使用 LDA 模型,但 LDA 模型在微博网络应用中有几个难以解决的问题:① 每个用户需单独建模,不同用户的主题数目难以确定,而且微博网络用户众多,主题庞杂;② LDA 对训练语料有较高的要求,训练语料越多,效果越好,而微博网络不同用户发表的微博数

量有较大的差别.

由于微博用语的不规范、新词的大量出现,采用传统的文本话题提取方法效果不够理想,近期的一些研究认为,有意义串在微博话题提取方面有较大的优势.有意义串是指具有统计意义、包含具体语义、能够独立灵活使用的语言单元.有意义串的识别可用于检索、分类领域以提高检索和分类的效率,也可以应用于频繁关键模式的抽取,以提取文本的分类或聚类特征等.比如,文献[27]提出的基于动量模型的微博突发话题检测方法,文献[28]提出的面向大规模微博消息流的突发话题检测,文献[29]提出的基于有意义串聚类的微博热点话题发现方法.本文借鉴了有意义串的思想从微博中提取兴趣(话题).

微博中有意义串的提取过程如下:(1) 对微博集 MB 中的每一篇微博 mb_i 进行分词,从分词结果中选出频率大于一定阈值 ξ_1 的词作为候选的有意义串集合 KW_1.为提取出现频率不高,但有实际意义的词,该阈值不宜过大.(2) 对 KW_1 中相邻且满足一定共现阈值 ξ_2 的词进行合并,得到候选的有意义串集合 KW_2.该过程需经过多次迭代统计,由原先的单个词逐渐合并得到越来越长的候选有意义串,直到最终没有可合并的串为止.(3) 计算 KW_2 中每个候选有意义串的上下文邻接类别来衡量其语用多样性,选取满足一定的上下文邻接类别阈值 ξ_3 的有意义串得到最终的有意义串集 KW.邻接类别越大,表明其使用越灵活,成为有意义串的概率就越大.

文献[27]经过实验发现,微博中能够作为话题的有意义串的上下文邻接类别的阈值为 2 或者 3 比较合理.一个有意义串的上下文邻接类别的阈值为 2 或者 3,说明该有意义串至少是出现了 2 次或者 3 次.参考这一结果,在过程(1)处的分词结果频率阈值、过程(2)处的分词合并阈值、过程(3)处的上下文邻接类别阈值分别设为 ξ_1=2,ξ_2=2,ξ_3=2.

用户发表的微博内容中,包含微博正文、话题(以符号"#...#"标识)和微博标签(常见的有以符号"[...]"标识的微博标题、类似于新闻的标题)等.文献[30]认为微博中的话题有较大的信息量,有意义串在话题及标题中的重要性更大,因此,对话题及标题中的有意义串的权重进行修正,方法如式(1)所示.

$$w(kw_i)=\begin{cases}\alpha_1\cdot w(kw_i), & kw_i\text{在微博的话题中}\\ \alpha_2\cdot w(kw_i), & kw_i\text{在微博的标题中}\end{cases} \tag{1}$$

权重系数 α_1,α_2 的最终取值通过实验确定,结果是,α_1 在[1.8,2.8],α_2 在[1.4,2.1]范围内比较合理.

用户在发表微博时有 3 种情况,包括原创、转发和评论的微博.用户发表的微博自然能够代表自己的兴趣,同时,虽然微博网络上有大量的博文,但用户只会就自己感兴趣的话题博文进行转发或评论.用户转发或者评论的微博同样也能表征自己的某些兴趣.因此,对评论及转发的微博内容中的有意义串的权重进行修正,方法如式(2)所示.

$$w(kw_i)=\begin{cases}\beta_1\cdot w(kw_i), & kw_i\text{在评论的微博中}\\ \beta_2\cdot w(kw_i), & kw_i\text{在转发的微博中}\end{cases} \tag{2}$$

参考文献[25,31],评论的权重系数设为 0.75,转发的权重系数设为 0.25.

KW 中的有意义串是个松散的集合,未能体现它们的关联.将 KW 中的有意义串依据互信息方法进行话题提取,即一个话题往往是由多个相互关联的有意义串构成的.

计算两个有意义串 kw_1,kw_2 互信息方法如下[32]:

$$MI(kw_1,kw_2)=\frac{f(kw_1,kw_2)}{f(kw_1)+f(kw_2)-f(kw_1,kw_2)} \tag{3}$$

其中,$f(kw_1,kw_2)$为在某个窗口范围内 kw_1 和 kw_2 共同出现的次数,由于微博比较短小,本文将窗口定义为每条微博范围内.

虽然微博网络包含了海量用户,其话题也涉及现实社会生活各个领域,但从统计学角度看,微博网络用户的话题符合幂律分布,即少量话题吸引了大量用户,而大量话题只被少数人关注.因此,可以提取用户的核心话题进行相似度计算,这一方面可以解决微博特征稀疏的问题,另一方面还可以减少计算的工作量.

话题重要度的计算如式(4)所示.

$$W(topic_i)=\sum_{j=1}^{n} w(kw_j)\cdot \log_2^{|T|} \tag{4}$$

式(4)在计算话题重要度时既考虑了话题中有意义串的出现次数,又考虑了话题的持续周期.

对兴趣点的权重参考最大值进行归一化,按照兴趣点权重降序排列,根据需求选取 m 个兴趣点作为用户的动态兴趣,记为 $UMBI=\{(topic_1,w_1,T_1),(topic_2,w_2,T_2),...,(topic_m,w_m,T_m)\}$.

2.4 UI-FP情形用户兴趣挖掘

对于用户背景及微博内容较少的情形,难以直接从该用户的信息中挖掘出其兴趣取向.用户的社交关系中包括关注和粉丝两类群体,还可以层层扩展,形成庞大的关联群体.文献[13]提及从用户朋友(直接或者间接有社交关系的用户)已有的数据中推理出用户的一些可能的背景信息,包括家乡、宗教、工作、性别、教育等,这与用户兴趣的挖掘有较大的不同.文献[17]发现,使用用户粉丝发表的微博提取用户兴趣,不如使用用户自己发表的微博提取用户兴趣的效果好.

本文提出的挖掘 UI-FP 情形用户兴趣与已有研究不同的是:① 在关联用户的选取上,在微博网络中用户间存在关注、粉丝和访客等几种情况,提出了基于关注的用户背景间接挖掘方法,通过第 4.3 节的实验验证了关注用户的使用效果较好;② 使用基于用户的"等级信息"和用户间交互强度的关注选取策略,提升了关注选取的有效性.

UI-FP 情形用户兴趣挖掘的核心问题转化为关注的选取策略,因为用户的关注可能很多,全部计算关注的兴趣不仅工作量大,而且依据二八法则,网络上 80%的内容是由 20%的用户创造的,选取出这部分用户已经很有代表性,可以避免其他用户杂乱兴趣点的干扰.

为用户 u_1 选取关注时,具体的步骤是:① 关注的等级,新浪微博的"等级信息"共分 24 个等级,能体现微博用户背景的丰富程度、发表微博的活跃程度等指标.将 24 个等级归一化为 0~1 之间的数值,关注 $follower_1$ 的等级得分记为 $L(follower_1)$.② 关注和 u_1 的交互性,用户间的交互性是指相互之间存在"@"、转发或者评论等行为,交互性能够反映用户的关联强度及兴趣点的重合性.用户间的交互强度使用他们之间的交互次数,将用户的交互次数进行归一化,关注 $follower_1$ 的交互得分记为 $I(follower_1)$.③ 计算关注 $follower_1$ 的最终得分,$S(follower_i)=I(follower_i)\times L(follower_i)$,按照得分降序排序,选取 m 个关注构成集合 FS.

在用户众多的背景信息中,简介、职业更多是用户个性化的描述,而标签则是同类众多用户的共性体现,因此标签更适合作为用户的背景兴趣,已有的相关研究也验证了使用用户的标签的优势所在,比如文献[12,14].UI-FP 情形的用户是指其背景和发表的微博信息量都较少.文献[33]在研究推荐系统的冷启动问题时,将评论信息条数少于 5 条的用户视为冷启动用户.文献[15]在研究用户的标签时,通过对新浪微博的 1.4 亿用户统计,发现标签数小于 5 的用户占用户总数的 93.8%.本文借鉴这些研究成果,当用户的标签个数及发表的微博条数都小于 5 的时候,认为此用户属于 UI-FP 情形.

用户 u_1 已有的标签集记为 $u_1.tag=\{tag_1,tag_2,\ldots,tag_e\}$,此时,$|u_1.tag|<5$.获取的关注 $follower_i\in FS$ 的标签集记为 $follower_i.tag=\{tag_1,tag_2,\ldots,tag_e\}$,计算 m 个关注的每个标签出现的次数,参考最大值对权重进行归一化.假设共需提取 n 个标签,按照权重降序排列,从关注提取的标签中选取 $n-|u_1.tag|$ 个标签,与用户已有的标签集合并,得到用户 u_1 的背景兴趣:$UI=\{Int_1,Int_2,\ldots,Int_n\}$,其中,用户 u_1 已有标签的权重设为 1.

3 基于背景和内容的用户兴趣相似性度量

用户 u_1,u_2 的静态兴趣相似度计算使用 Jaccard 方式,如式(5)所示.

$$UPISim(u_1.UPI,u_2.UPI)=\frac{|u_1.UPI\cap u_2.UPI|}{|u_1.UPI\cup u_2.UPI|} \tag{5}$$

用户 u_1,u_2 的动态兴趣中的两个兴趣点 Int_i,Int_j 的相似度计算如式(6)所示.

$$UMBISim(u_1.Int_i,u_2.Int_j)=\frac{Int_i.KW\cdot Int_j.KW}{\|Int_i.KW\|\cdot\|Int_j.KW\|}\cdot\frac{\min(Int_i.|T|,Int_j.|T|)}{\max(Int_i.|T|,Int_j.|T|)} \tag{6}$$

式(6)既考虑了兴趣点内容的相似度(余弦距离计算方法),又考虑了兴趣点的时间周期.在计算时间持续度时,并没有限定兴趣点的时间一致性问题,这是因为即使对同一个兴趣点,不同用户获取信息的时间不同,不同用户的认知能力不同,也都可能导致时间的偏差现象.

用户 u_1 和 u_2 的动态兴趣中的 m 个兴趣点的总相似度计算如式(7)所示.

$$UMBISim(u_1.UMBI,u_2.UMBI)=\sum_{i=1}^{m}\sum_{j=1}^{m}UMBISim(u_1.Int_i,u_2.Int_j) \tag{7}$$

对用户的静态兴趣相似度 $UPISim(u_1.UPI,u_2.UPI)$和动态兴趣相似度 $UMBISim(u_1.UMBI,u_2.UMBI)$进行整合,得到最终的用户兴趣相似度,如式(8)所示.

$$UISim(u_1,u_2)=\alpha\cdot UPISim(u_1.UPI,u_2.UPI)+(1-\alpha)\cdot UMBISim(u_1.UMBI,u_2.UMBI) \tag{8}$$

式(8)中,α是静态兴趣和动态兴趣权重的调节系数,$0\leqslant\alpha\leqslant1$,当$\alpha$=0 时,只使用用户的微博内容计算兴趣相似度;当$\alpha$=1 时,只使用用户的背景计算兴趣相似度.

4 实验及分析

4.1 实验数据

目前,还没有公开的用于微博用户兴趣挖掘、用户兴趣相似度计算的标准数据集.本文以新浪微博为例,选取了时尚、企业管理、教育、军事、文化这 5 个领域进行实验数据的采集、用户兴趣挖掘的分析.

在新浪微博搜索框中输入领域关键词进行检索,然后点击“找人”按钮,最多只能获取前 50 页用户,每页 20 个,共计 1 000 个用户.5 个领域使用的检索关键词及获取的用户数见表 1.

对表 1 获取的 6 684 个用户,进一步采集的信息包括:① 对 6 684 个用户进行 1 层关注、粉丝的扩展.新浪微博为防止他人获取用户的关注、粉丝进行恶意关注或广告骚扰,对非本人的关注、粉丝的访问量进行了限制,只能获取前 5 页内容,每页 20 个用户,关注和粉丝最多分别能获取 100 个用户,实际采集的用户总数为 714 472 个.② 采集 714 472 个用户的背景(简介、标签及职位)、发表的微博(包括原创、转发或评论),由于有些用户发表的微博过多,限制每个用户微博的发表时间是 2014 年 1 月 1 日~2014 年 12 月 31 日,共计采集微博 4 206 751 条.

Table 1 Keywords and number of obtained users in five fields

表 1 5 个领域的关键词及获取用户数

序号	领域	关键词	用户数
1	时尚	时尚潮流 美容护肤	1 000 1 000
2	企业管理	互联网高管 市场总监	66 1 000
3	教育	幼儿教育 高等教育	1 000 1 000
4	军事	歼 20	718
5	文化	谍战	900

4.2 用户兴趣相似度计算的准确率评价

(1) 实验方法

本部分共选用 7 种实验方法,分别介绍如下.

● 基于用户背景挖掘用户的兴趣,使用用户背景中的简介、标签及职位等信息,类似于文献[5]介绍的方法,简记为 RU-P.

● 基于用户背景挖掘用户的兴趣,使用用户背景中的标签信息,类似于文献[12,14]介绍的方法,由于背景中用户的兴趣点较少,新浪微博中用户的标签最多是 10 个,因此在进行用户兴趣相似度计算时,使用了静态兴趣

中的全部兴趣点,简记为 RU-PT.

● 本文提出的方法,基于用户背景挖掘用户的兴趣,使用用户背景中的标签信息,但如果用户背景较少(标签个数少于 5 个),则通过关注获取用户的背景,选取关注的数量为 30 个,简记为 RU-PT-F.

● 基于用户微博内容挖掘用户的兴趣,但不区分微博内容中的话题、标题,不区分微博是用户发表的,还是评论或者转发的,类似于文献[17,23]介绍的方法,微博中用户兴趣点的个数参考文献[25]所述,选取 50 个兴趣点,简记为 RU-MB.

● 基于用户微博内容挖掘用户的兴趣,但对用户转发或者评论的微博进行权重修正,类似于文献[25,31]介绍的方法,选取 50 个兴趣点,系数β_1,β_2 的取值参考文献[25,31],β_1=0.75,β_2=0.25,简记为 RU-MB-1W.

● 本文提出的方法,基于用户微博内容挖掘用户的兴趣,但对微博内容中的话题、标题进行权重修正,对用户转发或者评论的微博进行权重修正,用户兴趣点的个数为 50,系数α_1,α_2 的取值通过实验确定,α_1=2.5,α_2=1.5,简记为 RU-MB-2W.

● 本文提出的方法,基于用户背景和微博内容挖掘用户的兴趣,用户背景信息使用用户的标签,对微博内容中的话题、标题进行权重修正,对用户转发或者评论的微博进行权重修正,如果用户为 UI-FP 情形,则通过关注获取其兴趣.用户兴趣点为 50 个,选取关注的数量为 30 个,静态和动态兴趣整合时的权重系数α=0.6,简记为 RU-PMB.

(2) 实验结果

数据集中用户 u_1 的关注集记为 $u_1.follower$,作为标准答案,通过方法 RU-P 计算用户间的兴趣相似度选取出的关注集记为 $u_1.follower\text{-}RU\text{-}P$,令$|u_1.follower|=|u_1.follower\text{-}RU\text{-}P|$,方法 RU-P 选取的关注的准确率计算如式(9)所示.

$$RUA=\frac{\left|u_1.follower\cap u_1.follower\text{-}RU\text{-}P\right|}{\left|u_1.follower\cup u_1.follower\text{-}RU\text{-}P\right|} \tag{9}$$

其他 6 种方法计算获取关注准确率的方式与此类似.

5 个领域的 8 个检索关键词,每个随机选取 100 个用户,共计 800 个用户,7 种方法得到的平均准确率 RUA 见表 2.

Table 2 Average RUA of seven methods

表 2 7 种方法得到的平均 RUA

领域方向	方法						
	RU-P	RU-PT	RU-PT-F	RU-MB	RU-MB-1W	RU-MB-2W	RU-PMB
时尚潮流	0.436 2	0.479 1	0.583 3	0.425 3	0.457 1	0.460 3	0.634 5
美容护肤	0.326 5	0.502 3	0.577 1	0.393 8	0.412 2	0.422 7	0.583 6
互联网高管	0.416 9	0.503 3	0.578 1	0.411 6	0.430 3	0.441 1	0.624 4
市场总监	0.319 3	0.434 8	0.563 9	0.378 2	0.399 5	0.410 2	0.593 9
幼儿教育	0.409 7	0.477 9	0.568 8	0.420 4	0.431 1	0.440 8	0.613 4
高等教育	0.418 4	0.490 6	0.587 4	0.439	0.469 7	0.471	0.620 1
歼 20	0.420 5	0.502 1	0.582	0.424 4	0.475 1	0.486 7	0.648 3
谍战	0.441 4	0.517 2	0.559 3	0.412 7	0.453 4	0.480 2	0.630 7
平均值	0.398 6	0.488 4	0.575 0	0.413 2	0.441 1	0.451 6	0.618 6

由表 2 可见,7 种方法得到的平均准确率都不高,主要原因是即使计算出的用户兴趣相似度较高,该用户也并不一定成为其关注对象.RU-P,RU-PT 和 RU-PT-F 这 3 种方法都是从用户背景的角度出发获取用户的兴趣,由于对用户背景偏少的情况引入了关注的替代挖掘方法,RU-PT-F 的效果明显要好于 RU-P 和 RU-PT,这说明引入关注间接的挖掘用户兴趣的方式是有效的.方法 RU-PT 仅使用了用户背景中的标签,得到的结果比 RU-P 提高了 0.09 个点,说明使用用户背景中的标签挖掘用户的兴趣是有效的.RU-MB,RU-MB-1W 和 RU-MB-2W 这 3 种方法都是从用户发表微博的内容角度出发获取用户的兴趣,由于 RU-MB-2W 考虑了微博内容中的话题和标题,得到的结果要好于 RU-MB 和 RU-MB-1W.方法 RU-MB-2W 效果还不够明显的原因是,用户发表的很多微博内容中,包含话题、标题的并不是很多,对 400 多万条微博进行了统计,发现 31.6%的微博内容中包含话题,6.2%

的微博内容中包含标题.方法 RU-PT 和 RU-MB-2W 相比,基于背景的挖掘方法要好于基于内容的挖掘方法,主要原因是:一方面,微博中用户产生的文本信息中常常包含大量的口语、省略语、符号,这些文本的语义信息很难挖掘;另一方面,微博上有大量的非活跃用户,发表微博很少.方法 RU-PMB 得到的效果最为理想,验证了将微博用户的兴趣分为静态兴趣和动态兴趣,采用不同的策略分别计算的优势.

进一步使用 *P@n* 指标评价各种方法获取的用户兴趣是否真正相似.*P@n* 指标只关心计算用户兴趣相似度后得到的结果与用户兴趣是否相似,不考虑返回的用户之间的次序,人工评测起来容易实现.由于人工评测的工作量较大,所以 5 个领域的 8 个方向分别只选取了 10 个用户(共计 80 个用户),计算的结果只取前 10 个用户参与评测是否兴趣相似,即使用了 *P*@10 指标,结果见表 3.

由表 3 看以看出,5 个领域的 80 个用户的评测指标 *P*@10 得分普遍都在 0.74 以上,说明获取的用户间的兴趣相似度是比较高的.而表 2 反映出的情况是,挖掘出的用户真正属于关注的准确率却并不高,这说明用户在社交网络平台上构建的社交圈子还偏小,没能和很多兴趣相似的用户建立直接的朋友关系,这也是微博网络上信息推荐一直是热点研究问题的原因之一.方法 RU-P 和 RU-MB 的效果都偏差,同样是因为一些用户的背景或者发表的微博都较少,难以有效地计算用户的兴趣.方法 RU-PT-F 由于引入了关注获取用户的兴趣,效果得到明显的提升.方法 RU-MB-2W 由于进行了权重修正,效果也有所提升.RU-PMB 的效果是最好的,说明了融合用户的静态兴趣和动态兴趣计算用户兴趣的优势.

Table 3 Average *P*@10 of seven methods

表 3 7 种方法得到的平均 *P*@10

领域方向	方法						
	RU-P	RU-PT	RU-PT-F	RU-MB	RU-MB-1W	RU-MB-2W	RU-PMB
时尚潮流	0.7	0.85	0.9	0.7	0.8	0.85	0.95
美容护肤	0.7	0.85	0.85	0.75	0.75	0.85	1
互联网高管	0.65	0.8	0.85	0.7	0.85	0.9	1
市场总监	0.85	0.8	0.8	0.8	0.8	0.8	0.9
幼儿教育	0.75	0.8	0.9	0.75	0.75	0.75	0.95
高等教育	0.7	0.85	0.95	0.7	0.8	0.85	0.9
歼 20	0.75	0.9	0.9	0.8	0.8	0.8	0.9
谍战	0.8	0.85	0.9	0.75	0.85	0.85	1
平均值	0.74	0.84	0.88	0.74	0.80	0.83	0.95

4.3 关注和粉丝的选取对用户兴趣计算的影响

新浪微博网络对于一个用户,能够获取的关注和粉丝个数最多分别是 100 个,使用本文提出的方法 RU-PMB,关注和粉丝的取值分别从[10,20,30,…,100]进行实验.目的是:一方面检验选取关注还是粉丝间接挖掘用户的兴趣更有优势;另一方面,检验关注和粉丝数量的变化对准确率的影响.

由表 4 可以看出,选取关注挖掘用户的兴趣,得到的平均准确率 RUA 比选取粉丝的效果普遍偏好.同时,关注的个数选取为 30 时,效果已经比较理想,多选取关注不仅计算工作量大,而且效果并没有多大的提升.

Table 4 Average RUA with different followers and fans

表 4 不同关注和粉丝的个数得到的平均 RUA

个数	平均 RUA(关注)	平均 RUA(粉丝)
10	0.585	0.532 1
20	0.603 3	0.551 2
30	0.612 7	0.560 1
40	0.612 7	0.560 2
50	0.612 5	0.560 1
60	0.612 5	0.560 1
70	0.612 3	0.561 1
80	0.612 2	0.561 1
90	0.612 2	0.561 1
100	0.612 2	0.561 1

进一步地,验证从关注提取的用户背景与用户实际背景的差异.如第 2.4 节所述,实验时仅提取了关注的标签作为背景.从 5 个领域 800 个用户中选取标签比较丰富的用户,通过挖掘关注的标签与用户实际的标签比对来检验提取标签的效果.统计发现,在 800 个用户中,有 9.3%(74 个)的用户标签个数大于等于 5 个.用户 u_1 的实际标签集记为 $u_1.tag$,作为标准答案,从 u_1 的关注提取的标签集记为 $u_1.tag–F$,令$|u_1.tag–F|=|u_1.tag|$,两者之间的准确率计算如式(10)所示.

$$TA=\frac{\left|u_1.tag\bigcap u_1.tag-F\right|}{\left|u_1.tag\bigcup u_1.tag-F\right|} \tag{10}$$

5 个领域 74 个用户标签提取的准确率 TA 见表 5.

Table 5 Accuracy rate TA of 74 users' tags in five fields

表 5 5 个领域 74 个用户标签提取的准确率 TA

领域方向	TA
时尚潮流	0.631 7
美容护肤	0.653 2
互联网高管	0.673 9
市场总监	0.661 4
幼儿教育	0.745 2
高等教育	0.761 4
歼 20	0.655 1
谍战	0.674 3
平均 TA	0.682 0

由表 5 可以看出,通过关注提取的标签与用户实际的标签的准确率比较高,平均 TA 为 0.682.这说明通过用户的关注间接获取用户标签的方法是合理的,这为微博网络上大量用户缺少背景信息的描述(尤其是标签)的挖掘提供了途径.

4.4 静态和动态兴趣的权重对用户兴趣计算的影响

用户的最终兴趣是由静态兴趣 UPI 和动态兴趣 UMBI 整合而成,对权重系数α的取值从[0,0.1,0.2,…,1]进行实验,以检验权重系数对用户兴趣相似度计算的影响.选取不同的权重,使用方法 RU-PMB 得到的平均准确率 RUA 如图 3 所示.

由图 3 可以看出,权重系数α的取值范围在 0.5~0.7 的计算效果都比较理想,得到的平均准确率 RUA 普遍在 0.6 左右,因此建议α的取值范围在 0.5~0.7.当α=0 时,转化为基于微博内容计算用户兴趣相似度的方法 RU-MB,平均准确率 RUA 为 0.412 3;当α=1 时,转化为基于背景计算用户兴趣相似度的方法 RU-P-F,平均准确率 RUA 为 0.576 2.

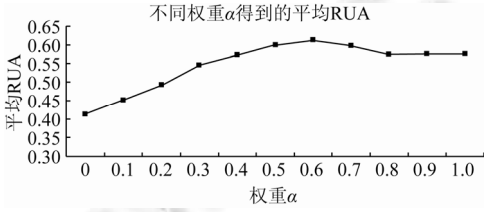


Fig.3 Average RUA with different weight α

图 3 不同权重α得到的平均 RUA

5 结束语

本文以微博用户的兴趣挖掘为出发点,在总结了已有研究方向的优点及不足的基础上,研究了微博网络的

表示模型、微博用户的静态兴趣及动态兴趣表示及挖掘方法,并给出了微博用户兴趣相似度的计算方法.

不考虑媒体的类型特征,仍然采用统一的用户兴趣表示模型,是不符合实际情况的.而且,挖掘用户的兴趣并不是最终目的,在用户兴趣的基础上,更多的是后期的用户关联度计算、社区划分、信息推荐等应用.本文提出的微博用户的动态兴趣表示模型由于引入了兴趣点的时间分布,可以为兴趣的动态分析提供基础.这种思想同样可以借鉴应用到事件、话题、用户之间的交互等动态特性相关的研究领域.

还需进一步提升的研究内容有:① 微博用户话题提取.由于微博用语过于灵活、不规范,且经常产生新的词语,导致微博话题的提取一直是研究的热点和难点.② 微博网络的海量信息处理.由于微博网络包含了大量的用户及用户生成的微博,有效地分析这类信息需要大数据处理相关技术的支持,包括分布式计算模型、NoSql型的数据存储及检索.③ 不同媒体类型的用户兴趣融合.用户在不同的社交媒体会生成诸多信息,为达到对用户的全面深入分析,需将来源于不同媒体的信息进行融合处理.

References:

[1] Liang YJ, Zheng XL, Zeng DD, Zhou XS, Leischow SJ, Chuang WY. Characterizing social interaction in tobacco-oriented social networks: An empirical analysis. Science Reports, 2015,5(16):1−11. [doi: 10.1038/srep10060]

[2] Wang CX, Guan XH, Qin T, Zhou YD. Modeling on opinion leader's influence in microblog message propagation and its application. Ruan Jian Xue Bao/Journal of Software, 2015,26(6):1473−1485 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4627.htm [doi: 10.13328/j.cnki.jos.004627]

[3] Guo L, Ma J, Chen ZM, Jiang HR. Incorporating item relations for social recommendation. Chinese Journal of Computers, 2014, 37(1):219−228 (in Chinese with English abstract).

[4] Wang XF, Tang L, Gao HJ, Liu H. Discovering overlapping groups in social media. In: Proc. of the 10th IEEE Int'l Conf. on Data Mining. IEEE Computer Society, 2010. 569−578. [doi: 10.1109/ICDM.2010.48]

[5] Diaby M, Viennet E, Launay T. Exploration of methodologies to improve job recommender systems on social networks. Social Network Analysis and Mining, 2014,4(227):1−17. [doi: 10.1007/s13278-014-0227-z]

[6] Ma H, Zhou D, Liu C, Lyu MR, King I. Recommender systems with social regularization. In: Proc. of the 4th ACM Int'l Conf. on Web Search and Data Mining (WSDM 2011). New York: ACM, 2011. 287−296. [doi: 10.1145/1935826.1935877]

[7] Kantor PB, Ricci F, Rokach L, Shapira B. Recommender Systems Handbook. New York: Springer-Verlag, 2009.

[8] Tsai WH, LinYT, Lee KR. Development of social-aware recommendation system using public preference mining and social influence analysis: A case study of landscape recommendation. Journal of Internet Technology, 2016,17(3):561−569. [doi: 10.6138/JIT.2016.17.3.20151110a]

[9] Cruz JD, Bothorel C, Poulet F. Entropy based community detection in augmented social networks. In: Proc. of the Int'l Conf. on Computational Aspects of Social Networks. 2011. 163−168. [doi: 10.1109/CASON.2011.6085937]

[10] Qi GJ, Aggarwal CC, Huang T. Community detection with edge content in social media networks. In: Proc. of the Int'l Conf. on Data Engineering. 2012. 534−545. [doi: 10.1109/ICDE.2012.77]

[11] Ghosh S, Sharma N, Benevenuto F, Ganguly N, Gummadi KP. Cognos: Crowdsourcing search for topic experts in microblogs. In: Proc. of the 35th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR). New York, 2012. 575−590. [doi: 10.1145/2348283.2348361]

[12] Liang C, Liu ZY, Sun MS. Expert finding for microblog misinformation identification. In: Proc. of the 24th ACL Int'l Conf. on Computational Linguistics. Mumbai, 2012. 703−712.

[13] Akcora CG, Carminati B, Ferrari E. User similarities on social networks. Social Network Analysis and Mining, 2013,3(3):475−495. [doi: 10.1007/s13278-012-0090-8]

[14] Xing QL, Liu L, Liu YQ, Zhang M, Ma SP. Study on user tags in Weibo. Ruan Jian Xue Bao/Journal of Software, 2015,26(7): 1626−1637 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4655.htm [doi: 10.13328/j.cnki.jos.004655]

[15] Wang X, Jia Y, Zhou B, Chen RH, Han Y. Interaction relation based user tag prediction in microblog site. Computer Engineering & Science, 2013,35(10):44−50 (in Chinese with English abstract).

[16] Ma YF, Zeng Y, Ren X, Zhong N. User interests modeling based on multi-source personal information fusion and semantic reasoning. In: Proc. of the 7th Int'l Conf. on Active Media Technology (AMT 2011). Berlin, Heidelberg: Springer-Verlag, 2011. 195−205. [doi: 10.1007/978-3-642-23620-4_23]

[17] Chen JL, Nairn R, Nelson L, Bernstein M, Chi EH. Short and tweet: Experiments on recommending content from information streams. In: Proc. of the SIGCHI Conf. on Human Factors in Computing Systems (CHI 2010). New York: ACM, 2010. 1185−1194. [doi: 10.1145/1753326.1753503]

[18] Weng JS, Lim EP, Jiang J, He Q. TwitterRank: Finding topic-sensitive influential Twitterers. In: Proc. of the 3rd ACM Int'l Conf. on Web Search and Data Mining. New York, 2010. 261−270. [doi: 10.1145/1718487.1718520]

[19] Zhao WX, in, Jiang J, Weng JS, He J, Lim EP, Yan HF, Li XM. Comparing Twitter and traditional media using topic models. In: Proc. of the 33rd European Conf. on Information Retrieval. Berlin, Heidelberg: Springer-Verlag, 2011. 338−349. [doi: 10.1007/978-3-642-20161-5_34]

[20] 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]

[21] 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]

[22] 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]

[23] Peng ZH, Sun L, Han XP, Shi B. Microblog user recommendation using learning to rank. Journal of Chinese Information Processing, 2013,27(4):96−102 (in Chinese with English abstract).

[24] Hong LJ, Davison BD. Empirical study of topic modeling in Twitter. In: Proc. of the 1st Workshop on Social Media Analytics. Washington, 2010. 80−88. [doi: 10.1145/1964858.1964870]

[25] 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]

[26] Zhao X, Chen RS, Fan K, Yan HF, Li XM. A novel burst-based text representation model for scalable event detection. In: Proc. of the 50th Annual Meeting of the Association for Computational Linguistics. 2012. 43−47.

[27] He M, Wang LH, Du P, Zhang J, Cheng XQ. Microblog hot topic detection based on meaningful string clustering. Journal on Communications, 2013,34(Z1):256−262 (in Chinese with English abstract).

[28] He M, Du P, Zhang J, Liu Y, Cheng XQ. Microblog bursty topic detection method based on momentum model. Journal of Computer Research and Development, 2015,52(5):1022−1028 (in Chinese with English abstract).

[29] Shen GW, Yang W, Wang W, Yu M. Burst topic detection oriented large-scale microblog streams. Journal of Computer Research and Development, 2015,52(2):512−521 (in Chinese with English abstract).

[30] Peng ZH, Sun L, Han XP, Chen B. Community hot statuses recommendation. Journal of Computer Research and Development, 2015,52(5):1014−1021 (in Chinese with English abstract).

[31] Xu ZM, Li D, Liu T, Li S, Wang G, Yuan SL. Measuring similarity between microblog users and its application. Chinese Journal of Computers, 2014,37(1):207−218 (in Chinese with English abstract).

[32] Zhang J, Gao JF, Zhou M. Extraction of Chinese compound words: An experimental study on a very large corpus. In: Proc. of the 2nd Workshop on Chinese Language Processing: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics. 2000. 132−139.

[33] Yang WS, Luo AM, Zhang MM. Trust-Circle based recommendation on user cold-start. Computer Science, 2013,40(11a):363−366 (in Chinese with English abstract).

附中文参考文献:

[2] 王晨旭,管晓宏,秦涛,周亚东.微博消息传播中意见领袖影响力建模研究.软件学报,2015,26(6):1473−1485. http://www.jos.org.cn/1000-9825/4627.htm [doi: 10.13328/j.cnki.jos.004627]

[3] 郭磊,马军,陈竹梅,姜浩然.一种结合推荐对象间关联关系的社会化推荐算法.计算机学报,2014,37(1):219−228.

[14] 邢千里,刘列,刘奕群,张敏,马少平.微博中用户标签的研究.软件学报,2015,26(7):1626−1637. http://www.jos.org.cn/1000-9825/4655.htm [doi: 10.13328/j.cnki.jos.004655]

[15] 汪祥,贾焰,周斌,陈儒华,韩毅.基于交互关系的微博用户标签预测.计算机工程与科学,2013,35(10):44−50.

[20] 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]

[23] 彭泽环,孙乐,韩先培,石贝.基于排序学习的微博用户推荐.中文信息学报,2013,27(4):96−102.

[25] 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]

[27] 贺敏,王丽宏,杜攀,张瑾,程学旗.基于有意义串聚类的微博热点话题发现方法.通信学报,2013,34(Z1):256−262.

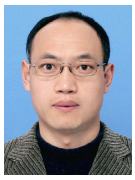
[28] 贺敏,杜攀,张瑾,刘悦,程学旗.基于动量模型的微博突发话题检测方法.计算机研究与发展,2015,52(5):1022−1028.

[29] 申国伟,杨武,王巍,于淼.面向大规模微博消息流的突发话题检测.计算机研究与发展,2015,52(2):512−521.

[30] 彭泽环,孙乐,韩先培,陈波.社区热点微博推荐研究.计算机研究与发展,2015,52(5):1014−1021.

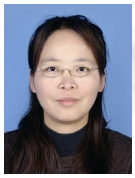
[31] 徐志明,李栋,刘挺,李生,王刚,袁树仑.微博用户的相似性度量及其应用.计算机学报,2014,37(1):207−218.

[33] 杨圩生,罗爱民,张萌萌.基于信任环的用户冷启动推荐.计算机科学,2013,40(11a):363−366.



仲兆满(1977－),男,江苏赣榆人,博士,副教授,主要研究领域为信息检索,文本信息挖掘,事件本体.

胡云(1978－),女,博士,副教授,CCF 专业会员,主要研究领域为数据挖掘,复杂网络.



管燕(1976－),女,讲师,主要研究领域为数据挖掘,模式识别.

李存华(1963－),男,博士,教授,主要研究领域为数据挖掘,人工智能,图像处理.