一种基于混合粒度的微博用户标签推荐模型

张 瑞 金志刚 王 颖
(天津大学电子信息工程学院 天津 300072)

摘 要 针对已有的标签推荐模型在实际微博场景运用中存在的多样性、相关性较差等不足,提出了一种基于混合粒度的标签推荐模型。将微博用户的可分析资源分解成由用户信息、标签和微博正文组成的混合粒度,在不同粒度上分别进行个人信息过滤及个性标签分析,从而计算用户标签的熵值与内联度和分类标注标签词汇,提取微博正文主题等,最终为用户推荐具有较强关联性的个性化标签。与一般LDA模型的对比实验证明,该模型可以有效解决新用户的冷启动、标签推荐的准确度等问题,同时保证了推荐的多样性。

关键词 社会化标签,混合粒度,主题提取,社交网络,多样性
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.4.039

Recommendation Model of Microblog User Tags Based on Hybrid Grain

ZHANG Rui JIN Zhi-gang WANG Ying
(School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract To avoid the shortcomings of existing tag-recommendation system, we proposed a recommendation model based on hybrid grain. The proposed model divides the resources to multiple grain including user information, tags and blog content. Then filtering personal information, analyzing personality tags, calculating user tag entropy and inline degree, classification words, and extracting topic of micro-blog can proceed based on the divided hybrid grain. Through the steps above, the proposed method can recommend associated personalized labels to users. By comparing the experimental results, it shows that the proposed model can efficiently deal with the cold boot problem, improve the accuracy of tag recommendation and insure the diversity of recommended labels.

Keywords Social tagging, Hybrid grain, Topic extraction, Social network, Diversity

中国社交网络用户数量日益增长,活跃用户数量已达2.2～2.5亿,每天发布在社交网络上的信息量以几何倍数增长,海量信息中蕴含的价值需要通过数据挖掘来获取。社交网络的海量数据挖掘成为了新兴的研究热点[1],但同时面临着巨大挑战。由于社交网络中的用户绝对数量巨大,用户关系挖掘存在着信息处理复杂度高、信息匹配准确率低等问题,因此引起了越来越多国内外研究者的关注。

近年来,微博成为了新兴的社交网络平台,其因内容简洁、即时互动以及快速传播等特性,成为当前互联网上进行新鲜观点获取、观点互动、意见发表以及消息传递的主流社会化媒体。用户之间的关联性已成为微博舆论发生、演进的重要路径和基础机制。目前大部分微博用户的研究都集中在用户核心影响力上,主要研究内容有核心用户的积聚性、集权性和圈群性[2],而对于普通用户之间的关联性及相似性缺乏相关研究。标签系统(Tagging System)是一种允许用户通过个人认知,自由地对自己或者其他用户进行标签标注的个性化评价系统,同时也是Web2.0时代下,对信息进行分类、组织、管理和搜索的重要手段[3]。国外的YouTube、Facebook、Instagram以及国内的QQ空间等一大批拥有标签系统的网站均获得了用户的欢迎。基于标签的个性化推荐,作为标签系统的组成部分,其可以为用户推荐更有针对性的信息,从而成为新的研究热点。

微博中的标签标注行为与一般标签标注不同,是由微博用户给自己进行标签标注,但是在实际行为中,用户由于对标签系统认知的不同以及本能地遵守"最省力原则"[4],因此无法获得良好体验。根据经过相关采集得到的2000余万个新浪微博用户个人信息可以看出,大部分用户并未给自己添加个性标签标注,如图1所示。因此,本文提出通过利用个人信息画像和热词抽取来对用户个性标签进行模型建立,找出具有相关规则的用户个性标签集合,从而为微博用户进行更加准确的个人标签推荐。

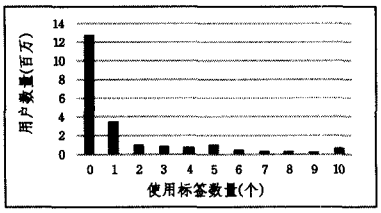


图1 2000余万个用户的标签使用数统计

到稿日期:2015-04-21 返修日期:2015-08-05 本文受国家自然科学基金资助项目(61201179)资助。

张 瑞(1986－),男,博士生,主要研究方向为数据挖掘、舆情分析;金志刚(1972－),男,教授,博士生导师,主要研究方向为无线网络、物联网、网络安全、网络管理;王 颖(1977－),女,博士,主要研究方向为网络安全。

1 背景知识及相关工作

微博系统中的用户个性标签系统是一个由用户和标签组成的双元关联系统。以新浪微博为例，用户在注册时以及使用中，都可以为自己添加个人标签（见图 2），用以描述有关自己的职业、兴趣等，通过个性标签，用户可以发现更多的同类人群，形成超越现实社会关系的新社交关系，进而形成不同的社交圈落。

图 2 新浪微博个人标签信息

现有的微博个人标签推荐方式存在一些不足，以用户量最多的新浪微博为例，新浪微博个人标签推荐是基于随机性的推荐，用户每次刷新可以重新获取 10 个推荐标签。这种机制虽然可以保证标签词汇不会过于集中在某几个热门词汇上，但是效率低，用户经常需要刷新数下乃至数十下才可以获取恰当的标签，这种机制下的标签体验较差，导致个人标签这一重要的社会属性并没有得到用户的重视。产生此现象有两方面原因：一方面是由于个人标签的推荐没有激发用户选择兴趣，也就是遇到了推荐算法中“冷启动”的问题[5]，无法有效进行推荐；另一方面是个人标签设置完成后，用户并没有针对这一属性进行相关使用。另外，在用户选择的标签中，音乐、旅游、电影等一般词汇出现频率非常高，如果单纯利用关联规则进行推荐词语的规则查找，非常容易出现推荐的同质化现象，使得用户标签这一个选择行为失去多样性。

在标签推荐系统的研究中，主要存在的重点、难点包括：(1)特征提取；(2)模型过拟合；(3)新用户加入；(4)新对象加入；(5)数据稀疏问题等[6]。针对这些问题，不少学者都提出了自己的解决方法，例如为了解决特征提取和模型拟合等问题，Guan 等人结合用户偏好度和文档相关度，提出了多类别相关对象排序法来为用户推荐标签[7]；Zhang 等人结合语言模型提出了 ACT(Author-Conference-Topic)模型，提高了标签推荐的准确率[8]；Xu 等人结合好友关系以及 LDA(Latent Dirichlet Allocation)主题模型，提出了改进的基于好友关系约束的 RBLDA 模型进行微博个性标签推荐[9]。为了解决新用户加入和新对象加入等问题，Chen 等人以微博用户的关注者或者粉丝作为标签源进行文本分类以进行用户标签推荐[10]；Hu 等人提出了利用拓扑关系来解决标签推荐的冷启动问题[11]。为了解决数据稀疏问题，Cai 等人提出了一种结合用户标签和协同过滤算法的混合推荐方法[12]；Liao 等人采用了张量分解方法来进行标签推荐[13]；王莎等人基于标签提供了一种微博人脉网络挖掘算法[14]。但文献[7,8,11-13]没有针对微博个性标签这一特定环境进行考虑，缺乏适用性；而文献[9,10]没有对数据稀疏问题提出合适的解决方案；文献[14]利用了微博中标签的特点进行指定主体的用户相似性分析，但是没有进一步提出如何对混合主题的用户进行分析推荐。

通过分析发现，以上方法都是针对传统物品推荐系统进行改进，没有重视微博标签具有口语化、碎片化，以及无评价性等特点，已有传统方法无法完全适用于微博中的标签推荐。

因此，本文提出一种基于混合粒度主题模型的微博个性化标签推荐模型 HG-LDA(Hybrid-Grain LDA)，通过对微博用户在不同粒度上的分析处理，改善传统 LDA 的性能，进而优化个人标签的添加过程，提高个人标签这一社交属性的使用率和实用性，同时解决了推荐的冷启动问题。

2 HG-LDA 模型

HG-LDA 模型包括微博数据获取、混合粒度组成、粒度组成、粒度间关系分析。由 HG-LDA 模型构建的微博用户标签推荐平台如图 3 所示。

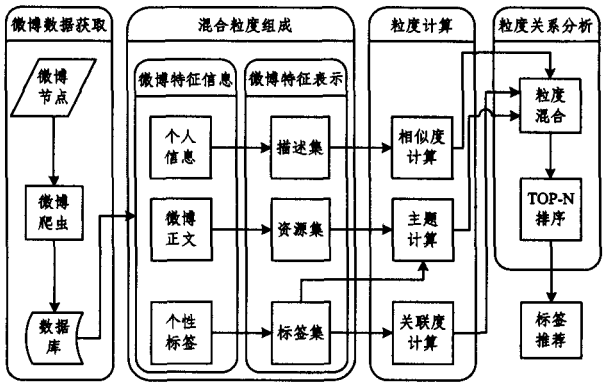


图 3 混合粒度的微博用户标签推荐模型

HG-LDA 模型的微博数据获取主要包括两个部分：微博用户信息数据获取和微博语料获取。本文采用中国爬盟的新浪微博爬虫，构建微博用户信息库。该爬虫是由清华大学智能技术与系统国家重点实验室信息检索组制作，通过众包方式爬取微博数据。经过近两个月的采集与整理，共计获得 22399815 条用户数据，并存入微博用户信息数据库。本文采用由自然语言处理与信息检索共享平台(http://www.nlpir.org)提供的 NLPIR 微博语料库。该微博语料库涵盖了 23 万条微博内容语料。

对于一个给定用户(user)，其用户信息包含 3 种有效信息，分别是个人背景信息、历史微博正文、个性标签集合。因此，可以将任一用户表示为{$background(user)$, $text(user)$, $tag(user)$}的混合粒度问题，即可以将这 3 个粒度的信息处理表述为{描述集，资源集，标签集}的模型分析。

HG-LDA 模型中的混合粒度组成设计和在 HG-LDA 中解决冷启动问题是本文的重点工作，下面将进行详细介绍。

2.1 描述集模型

在推荐系统中，对于一个用户而言，总是存在大量的没有经过该用户评价或者查看的数据，而且这类数据常常比该用户已评价的数据量更大。在微博环境下，研究者更加关注用户的博文、图片和评论等内容，而对用户的个人信息、个性化标签等隐性数据存在一定的忽视。用户之间由于选择的差异性非常大会造成数据稀疏的情况，即任意两个用户的评分差别都非常大。本文使用微博个人资料中提供的用户所在地、性别、出生日期(星座、年龄)、血型、教育信息、职业单位等个人信息作为描述集计算的依据，如图 4 所示。

先对用户的个人信息进行相似度计算，可以避免由于新加入用户没有历史数据而无法使用推荐模型进行计算的问题，即冷启动问题。设用户个人信息为字符串集合 $p(user)=$

{*location*, *sex*, *age*, *constellation*, *blood*, *education*, *career*}，将该集合组合成长文本 *ProfileString*(*user*)，简称 *ps*(*u*)，对其进行字符串相似度分析。本文采用基于最长公共子串的文本比较算法——LCS[14]。

根据本文实验需要，将计算得到的 LCS 长度值与字符串的最大长度进行比较，得出[0,1]之间的比例值，该值越大则两个字符串间的相似度越高。即：

$$p(ps(A),ps(B))=\frac{LCS(ps(A),ps(B))}{MaxLength(ps(A),ps(B))} \tag{1}$$

若计算得到的结果有一个匹配度值大于阈值 x，则认为该用户满足相似度条件，x 的取值将在后续的实验中进行分析。

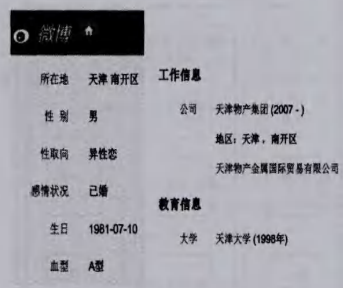


图 4 新浪微博个人资料

2.2 标签集模型

N-Gram 是为了弥补分词系统局限性而采用的一种效果较好的方法，本文也采用了该方法。但由于该方法的一般实现是：先抽取出 1-Gram，然后在其基础之上进行迭代扩展短语，因此有较大可能漏掉一些比较重要的短语。比如“奥巴马”，分词系统会将其拆解为“奥”、“巴”、“马” 3 个词汇，丢失原有的信息内容。对此，我们通过将所有小于等于 N 的 N-Gram 全部抽取出来进行一次遍历筛选来避免上述问题。如一个句子由 n 个词汇组成，即 $S=\{w_1,w_2,\cdots,w_n\}$，由图 1 可知大量用户一般选择 5 个以下的个性化标签，因此本实验中，选择 $N=4$，则会得到：

$$W=\sum_{i}^{i+1}w_i,i\in[1,N-1] \tag{2}$$

W 为 S 语句中每两个相邻词的集合，判断 W 是否有意义，可以由两点确定：1)W 的边界是否清晰，2)W 内部是否结合紧密。短语组合的边界是否清晰通过条件熵与联合熵共同判断，当条件熵和联合熵分别大于指定阈值时，认为短语的边界清晰；内部是否紧密通过内联度 *NLD* 判断，通过 *NLD* 的值，可以得到语料中标签短语中每个组成词汇的产生顺序和相互关系，即可以获得短语组合的内部紧密度。

若得到符合以上标准的 W，且 W 由不少于一个的词汇 w 组成，就可以认为该 W 表达了两个词汇之间的关联程度，即在标签推荐中可以优先推荐 W 中的词汇作为参考。

在从标签集粒度上进行推荐时，可以将内联度大于阈值 N 的项作为较好的连续性关联推荐提供给用户。例如通过实验发现，“投资”和“理财”两个标签的关联度很高，那么在用户选择其中一个标签时，可以将另外一个标签作为推荐。

为了保证进一步提高所推荐标签与用户资源集取向一致性，对于得到的标签集再进行基于朴素贝叶斯准则的词袋模型公式（如式(3)所示）训练，以为资源集的主题词汇分类提供可用词袋，将训练结果加入到 2.3 节中的资源集训练模型中。

$$p(x|y_i)p(y_i)=p(y_i)\prod_{j=1}^{m}p(a_j|y_i) \tag{3}$$

其中，x 为待分类标签集合，m 为该集合个数，a 为该集合的某一项标签。y 为有类别集合，即 x 为已知分类 y 的待分类项集合。通过以下 3 个步骤可以得出标签的分类集合。

1)准备工作：根据具体情况确定特征属性，对每个特征属性进行适当划分，由人工对一部分待分类项进行分类，形成训练样本集合。

2)分类器训练：计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估计。

3)应用分类器：使用分类器对待分类项进行分类。

2.3 资源集模型

LDA[16]是一种无监督学习的主题概率生成模型，输入是文档集合和主题个数，输出是以概率分布的形式呈现的主题，常用于主题建模、文本分类、观点挖掘等多个领域。LDA 假定了一个前提：文档相当于一个词袋（bag-of-words），袋子中的词是独立可交换的，没有语法结构和顺序。其基本思想是：每个文档(Document)由多个主题(Topic)构成，每个主题都由对应的多个词(Word)来描述。其模型描述如图 5 所示。

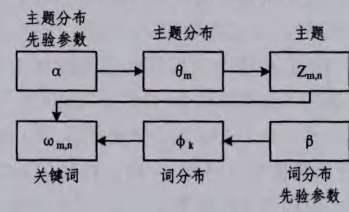


图 5 LDA 主题模型

其中，$m\in[1,M]$，$n\in[1,N_m]$，$k\in[1,K]$。该模型中 K 为主题个数，M 为文档总数，N_m 为第 m 个文档的单词总数。β 是每个主题下词的多项分布的狄克雷先验参数，α 是每个文档下主题的多项分布的狄克雷先验参数。$Z_{m,n}$ 是第 m 个文档中的第 n 个词。θ_m 和 φ_k 分别表示第 m 个文档下的主题分布和第 k 个主题下词的分布。对于给定的文档集合，$\omega_{m,n}$ 是可以观察到的已知变量，α 和 β 是根据经验给定的先验参数，其他的变量 θ_m 和 φ_k 都是未知的隐含变量，需要根据观察到的变量来进行学习估计。在文献[16]中，对于主题个数为 K、单词个数为 V 的文档环境，采用 Gibbs 采样方法可以得到 θ_m 和 φ_k 的计算公式：

$$\theta_{m,k}=\frac{n_m^{(k)}+\alpha_k}{\sum_{k=1}^{K}n_m^{(k)}+\alpha_k} \tag{4}$$

$$\varphi_{k,t}=\frac{n_k^{(t)}+\beta_t}{\sum_{t=1}^{V}n_k^{(t)}+\beta_t} \tag{5}$$

由此可以给出计算当前词的主题概率的公式：

$$(z_i=k\mid z_{-i},w)=\frac{n_{k,-i}^{(t)}+\beta_t}{\sum_{t=1}^{V}n_{k,-i}^{(t)}+\beta_t}\cdot\frac{n_{m,-i}^{(k)}+\alpha_k}{[\sum_{k=1}^{K}n_m^{(k)}+\alpha_k]-1}\propto\frac{n_{k,-i}^{(t)}+\beta_t}{\sum_{t=1}^{V}n_{k,-i}^{(t)}+\beta_t}(n_{m,-i}^{(k)}+\alpha_k) \tag{6}$$

模型中 LDA 以得到分类结果的标签集与微博语料库为输入，生成主题和词汇的分布。最后将生成的分布结果应用于资源集，得到推荐结果。对于 LDA 中主题分布和词汇分布

两个先验参数 α 和 β 的选择，通过在标签集上进行交叉验证获得。

2.4 HG-LDA 模型算法

结合微博用户标签的特点，首先对已有用户进行描述集相似度筛选，减少噪声数据量；然后对标签数据进行预处理，筛选出符合可以组成短语要求的用户标签集合，然后对该集合进行短语语料库的生成，遍历该短语语料库，筛选出符合条件熵、联合熵以及内联度要求的短语集合，同时得到标签分类结果；将得到的标签集分类结果与语料库进行主题与词汇的分布生成，将训练结果与资源集拟合，得到最后的推荐标签集输出。

2.5 冷启动解决

为了解决微博用户冷启动的情况，先根据用户填写的个人信息，如地区、年龄、星座等，在已有用户集合中进行相似度计算，可以筛选出有相似性的用户集合；对此集合中的用户标签进行抽取及关联度计算，得到相关联标签集；再对该相关联标签集进行主题分析，这里针对一般 LDA 中需要解决的主题分布先验参数和词分布先验参数不易取得的问题，使用相关联标签集进行训练，调试出合适的先验参数；再将得到的主题分布与微博语料库进行词分布训练，得到标签所涵盖的主题下的词袋，以此作为训练集结果用于相似性用户集合中；对其集合所包括的博文集合进行主题挖掘，得到相关性较大的主题性标签推荐，可以将该主题下涵盖的词汇作为标签推荐给用户；其后可将用户不断增加的博文也引入模型当中进一步进行训练，使得标签推荐随着用户兴趣的进一步发现而调整推荐内容，得到更加准确的结果。

3 实验

本文按照图 6 所示过程进行实验，实验所使用的运行系统配置为 2.4GHz i7 处理器、16GB 内存、64 位 Windows 8.1 操作系统以及 Java1.6SE 版本环境。分别采用一般的 LDA 算法[18]和 HG-LDA 算法进行实验。两种方法均使用 Java 语言实现，其中 LDA 算法采用版本进行对比测试。性能测试采用推荐系统常用的准确度(Precision)、召回率(Recall)以及 F-度量(F-measure)进行比较分析。

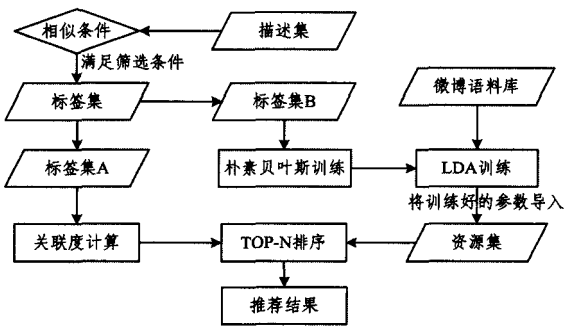


图 6 基于混合粒度的标签推荐流程

3.1 实验数据集及预处理分析

使用中国爬盟的新浪微博爬虫采集本实验的数据集，通过近两个月的采集与整理，共计获得 22399815 条用户数据，筛查掉 0 标签用户后，共计 9659120 条用户信息，从此实验数据集中随机抽取标签使用数量在 4 个以上的用户共计 18 万余个。经过数据预处理得到 18 万余个用户标签集 D_1，根据 Microsoft 最新的数据挖掘定义[19]，按照 7:3 的比例，取 12.6 万为系统训练样本，其它 5.4 万为系统测试样本，分别进行以下训练和测试。根据对训练样本的个人信息描述集相似度统计，得到如图 7 所示结果。为了使得筛选后可以保证有相当基数的测试范围量，选择阈值 x 为 0.3，将符合要求的用户作为相似用户集 D_2，进行个性化标签推荐实验。

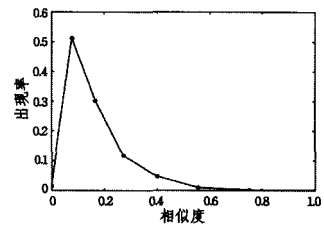
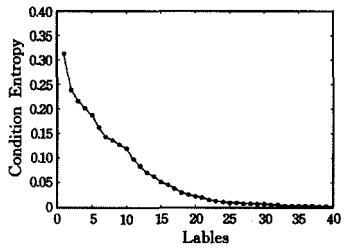


图 7 用户相似度分布

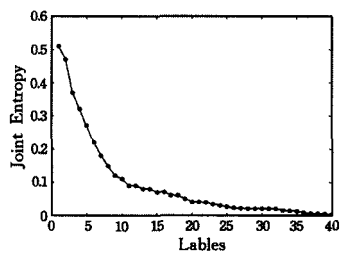
3.2 混合粒度标签推荐实验

在上述数据集中，将 3.1 节中已经生成的标签短语文档 D_3 导入到一般 LDA 和 HG-LDA 中进行模型训练和计算。为了便于数字精度简化，将所得到的熵值等均扩大 10 倍，使读数更加方便。

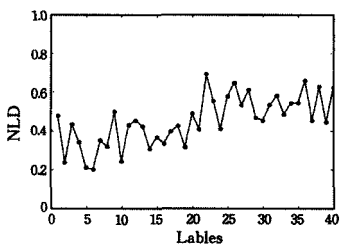
在标签库中选择出现率最大的 500 个标签作为指定词汇进行熵值计算，对所得熵值取平均后，可以发现与指定词汇最相关的 40 个标签有如图 8 所示的结果。



(a)条件熵均值



(b)联合熵均值



(c)内联度均值

图 8 阈值均值

因为微博个性标签推荐每一次备选个数为 10，根据图 8 所示，将条件熵阈值 C 设置为 0.01，联合熵阈值 J 设置为 0.05，内敛度阈值 N 设置为 0.25，可以保证备选标签总数保持在 20 个以上，使得用户的选择可以保证多样性。下面分别以“投资”、“教育”这两个标签为例进行计算，得到表 1 所列结果。

表 1　RHPEP 计算结果(前 10 个)

投资	理财	数码	咨询	健康	娱乐
条件熵	0.31	0.23	0.21	0.20	0.18
联合熵	0.46	0.36	0.34	0.33	0.30
内联度	0.31	0.65	0.45	0.28	0.37
投资	汽车	时尚	文艺	80 后	体育
条件熵	0.13	0.11	0.07	0.05	0.04
联合熵	0.23	0.20	0.14	0.10	0.08
内联度	0.27	0.42	0.65	0.71	0.61
教育	就业	学生	90 后	娱乐	体育
条件熵	0.65	0.56	0.45	0.36	0.31
联合熵	0.78	0.68	0.57	0.48	0.43
内联度	0.47	0.41	0.55	0.64	0.72
教育	新闻	理财	生活	健康	时尚
条件熵	0.25	0.20	0.16	0.11	0.07
联合熵	0.35	0.30	0.24	0.18	0.12
内联度	0.42	0.81	0.92	0.71	0.74

对于已得到的相似用户集，对其标签进行数量汇总排序，对出现次数最多的 500 个标签进行人工分类，设定主题数为 50 个，可得到微博用户的标签主题集和词汇集关系，表 2 为统计结果示例。

表 2　微博用户标签分类

标签	类别
旅行、美食、听歌、旅游、音乐、体育等	兴趣
刘德华、五月天、罗志祥、郭德纲等	明星
内向、沉默、活泼、真诚等	性格
教师、学生、码农、医生、职员等	职业
…	…

将该分类结果用于所有标签，可以由朴素贝叶斯方法计算出全部标签的分类集合，即可以作为 HG-LDA 模型的主题集和单词集，将该结果加入到微博语料库中进行进一步的训练。

对于系统性能的评测，本文采取准确率、召回率、F 度量 3 个指标来对标签推荐性能进行评价，具体定义如下：

$$recall=\frac{\text{正确抽取的标签}}{\text{标签总数}} \tag{7}$$

$$precision=\frac{\text{正确抽取的标签}}{\text{抽取标签总数}} \tag{8}$$

$$F=\frac{2*precision*recall}{precision+recall} \tag{9}$$

对于 LDA 和 HG-LDA 中的 α 和 β 参数，由文献[16]可知，由一般经验，α 参数可以设定为 2，β 参数设定为 0.5。这里结合实际微博场景对 α 和 β 参数进行重新拟合，待训练的 α 和 β 参数如表 3 所列，可以得到图 9 所示结果。

表 3　待训练参数

集合	1	2	3	4	5
α	1.7	1.7	1.7	1.8	1.8
β	0.4	0.5	0.6	0.4	0.5
集合	6	7	8	9	10
α	1.8	1.9	1.9	1.9	2.0
β	0.6	0.4	0.5	0.6	0.4
集合	11	12	13	14	15
α	2.0	2.0	2.1	2.1	2.1
β	0.5	0.6	0.4	0.5	0.6

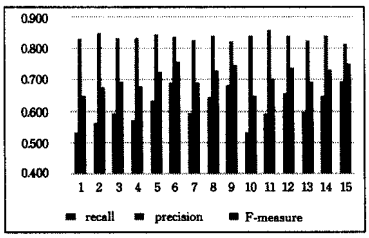


图 9　HG-LDA 参数估计结果

可见看到第 6 组集合的 F-Measure 是最高的，故采用第 6 组进行推荐计算，即 α 和 β 分别为 1.8 和 0.6。在此参数下，将使用 HG-LDA 算法在 5.4 余万大小的测试集中得到的推荐标签与由 LDA 算法得到的标签相比较，对比得到表 4 所列结果。

表 4　给定标签的推荐结果对比

给定标签	算法模型	结果标签词(前 10 个)
教育	LDA	音乐；旅行；旅游；美食；时尚；90 后；培训；就业；电影；学习
	HG-LDA	就业；学生；90 后；娱乐；体育；新闻资讯；理财；生活；健康；时尚
投资	LDA	音乐；美食；电影；旅行；80 后；90 后；理财；数码；健康；新闻
	HG-LDA	理财；数码；资讯；健康；娱乐；汽车；时尚；文艺；80 后；体育

可以明显看出，相对于 LDA 模型中的普遍化推荐，如音乐、美食、旅行等出现频率非常高的词汇，HG-LDA 模型可以根据给定标签，推荐更加符合此标签主题的词汇供用户选择，使得用户的使用体验更好。

在 α 取 1.8、β 取 0.6 的情况下，根据测试结果可以分别计算 HG-LDA 推荐算法的准确度、召回率以及 F-度量，与 LDA 算法对比的结果如表 5 所列。

表 5　标签推荐性能评价

实验模型	recall(%)	precision(%)	F(%)
LDA	81.31	41.27	54.75
HG-LDA	83.63	69.12	75.69

从表 5 可以看出在微博标签环境下，相比于一般 LDA 算法，HG-LDA 模型可以获得一定的性能提高，尤其是在准确率和总体性能(F 值)方面提高幅度较大。相对于 LDA 算法，HG-LDA 算法推荐的准确度提升了 27.85%，F-度量提升了 20.94%。因此，通过以上实验，可以看到 HG-LDA 在推荐多样性、相关性、推荐性能上都优于单纯地使用一般 LDA 算法。

结束语　本文针对微博用户标签使用率不高的现状，通过深入分析微博标签的实际特点，提出了一种基于混合粒度的个性化标签推荐模型算法，这一模型可以由发现用户相似性来缩小标签推荐范围及降低运算量，并且可以表示出标签之间的连接程度。通过使用来自真实应用的数据集验证该方法，并与一般 LDA 模型算法进行对比，实验结果表明该方法推荐的标签具有多样性丰富、结合程度更加准确的特点，而且避免了传统推荐系统中存在的数据稀疏性和冷启动问题。下一步的研究工作包括设定更精确的阈值以及冗余和相似词条的筛选，对个人信息的相似度计算中引入权值，以进一步提高算法的准确度和精确度，以满足其他信息资源推荐中的应用；同时还可以将该模型应用于基于标签的相似兴趣用户推荐或相似资源推荐中。

参 考 文 献

[1]　Di L, Du Y P. Application of LDA Model in Microblog User Recommendation[J]. Computer Engineering, 2014(5): 1-6, 11 (in Chinese)
邸亮，杜永萍. LDA 模型在微博用户推荐中的应用[J]. 计算机工程，2014(5)：1-6，11

(下转第 230 页)

sequence labeling[C]// Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP). 2013

[9] Mansur M, Pei W, Chang B. Feature-based Neural Language Model and Chinese Word Segmentation[C]//International Joint Conference on Natural Language Processing. 2013:1271-1277

[10] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. The Journal of Machine Learning Research, 2003,3:1137-1155

[11] Mikolov T, Kombrink S, Burget L, et al. Extensions of recurrent neural network language model[C]// 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011:5528-5531

[12] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011, 12:2493-2537

[13] Zheng X, Chen H, Xu T. Deep Learning for Chinese Word Segmentation and POS Tagging[C]//EMNLP. 2013:647-657

[14] Pei W, Ge T, Baobao C. Maxmargin tensor neural network for chinese word segmentation[C]//Proceedings of ACL. 2014

[15] Liu J S. Monte Carlo strategies in scientific computing[M]. Springer Science & Business Media, 2008

[16] Hinton G. Training products of experts by minimizing contrastive divergence[J]. Neural Computation, 2002, 14(8):1771-1800

[17] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. Neural Computation, 2014, 14:1771-1800

[18] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]// Advances in Neural Information Processing Systems. 2013: 3111-3119

[19] Mikolov T, Yih W, Zweig G. Linguistic Regularities in Continuous Space Word Representations[C]// HLT-NAACL. 2013: 746-751

(上接第 196 页)

[2] Ding Z Y, Jia Y, Zhou B, et al. Survey of Influence Analysis for Social Networks[J]. Computer Science, 2014, 41(1): 48-53(in Chinese)
丁兆云,贾焰,周斌,等. 社交网络影响力研究综述[J]. 计算机科学,2014,41(1):48-53

[3] Denning P J. Computing is a natural science[J]. Communications of the ACM, 2007, 50(7):13-18

[4] Jiang W Q. Zipf and the Principle of Least Effort[J]. Tongji University Journal(Social Science Section), 2005, 16(1):87-95(in Chinese)
姜望琪. Zipf 与省力原则[J]. 同济大学学报(社会科学版), 2005,16(1):87-95

[5] Sun D T, He T, Zhang F H. Survey of Cold-start Problem in Collaborative Filtering Recommender System[J]. Computer and Modernization, 2012(5):59-63(in Chinese)
孙冬婷,何涛,张福海. 推荐系统中的冷启动问题研究综述[J]. 计算机与现代化,2012(5):59-63

[6] Zhang Z K, Zhou T, Zhang Y C. Tag-aware recommender systems: a state-of-the-art survey[J]. Journal of Computer Science and Technology, 2011, 26(5):767-777

[7] Guan Z, Bu J, Mei Q. Personalized tag recommendation using graph-based ranking on multi-type interrelated objects[C]// Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2009:540-547

[8] Zhang N, Zhang Y, Tang J. A tag recommendation system based on contents[C]// Proceedings of the ECML PKDD Discovery Challenge Workshop, 2009(DC09). 2009:285

[9] Xu B, Yang D, Zhang Y, et al. Relationship Bind Topic Model Toward Tag Recommendation for Micro-Blog Users[J]. Journal of Frontiers of Computer Science and Technology, 2014(3): 288-295(in Chinese)
徐彬,杨丹,张昱,等. 面向微博用户标签推荐的关系约束主题模型[J]. 计算机科学与探索,2014(3):288-295

[10] Chen Y, Lin L, Sun C J, et al. A Tag Recommendation Method for Microblog Users[J]. Intelligent Computer and Applications, 2011(5):21-26(in Chinese)
陈渊,林磊,孙承杰,等. 一种面向微博用户的标签推荐方法[J]. 智能计算机与应用,2011(5):21-26

[11] Hu J, Wang B, Liu Y. Personalized tag recommendation using social influence[J]. Journal of Computer Science and Technology, 2012, 27(3):527-540

[12] Cai M S, Li X M, Yin Y T. Hybrid top-N recommendation method based on social user tag [J]. Application Research of Computers, 2013(5):1309-1311, 1344(in Chinese)
蔡孟松,李学明,尹衍腾. 基于社交用户标签的混合 top-N 推荐方法[J]. 计算机应用研究,2013(5):1309-1311,1344

[13] Liao Z F, Wang C Q, Li X Q, et al. Tag Recommendation and New User Tag Recommendation Algorithms Based on Tensor Decomposition [J]. Journal of Chinese Computer Systems, 2013, 34(11):2472-2476(in Chinese)
廖志芳,王超群,李小庆,等. 张量分解的标签推荐及新用户标签推荐算法[J]. 小型微型计算机系统,2013,34(11):2472-2476

[14] Wang S, Zhang L M. Mining Algorithm and Structural Analysis of Microblog Interpersonal Relationship Network Based on Tag [J]. Computer Engineering, 2014, 40(5):7-11(in Chinese)
王莎,张连明. 基于标签的微博人脉网络挖掘算法和结构分析[J]. 计算机工程,2014,40(5):7-11

[15] Atallah M J. Algorithms and Theory of Computation Handbook [M]. CRC Press, 1998

[16] Chen L F, Mark-Liao H Y, Ko M T, et al. A new LDA-based face recognition system which can solve the small sample size problem[J]. Pattern Recognition, 2000, 33:1713-1726

[17] Heinrich G. Parameter estimation for text analysis[R]. 2004

[18] Liu Yang, Qiu Ming-hui, Gottipati S, et al. CQARank: Jointly Model Topics and Expertise in Community Question Answering [C]// Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013). 2013:99-108

[19] 定型数据集和测试数据集[OL]. http://msdn.microsoft.com/zh-cn/library/bb895173.aspx