不同粒度标签推荐算法的比较研究 *

靳延安[1,2], 李玉华[1], 刘行军[2]
(1. 华中科技大学 计算机学院, 武汉 430074; 2. 湖北经济学院 信息管理学院, 武汉 430205)

摘 要: 针对社会标签系统中不同粒度的特征在表示文档时具有不同的描述能力这一特性,提出从词粒度和话题粒度来推荐社会标签以提高标签推荐的准确度。提出使用统计语言模型(词粒度)和隐含话题模型(话题粒度)分别建模文档的描述集和标签集,首先使用单个模型进行标签推荐,然后融合不同的特征粒度进行标签推荐。实验结果表明:就单一方法讲,基于统计语言模型的推荐性能要比基于话题粒度模型的推荐性能好;基于两种方法的混合方法的性能要好于没有混合的基于话题的单个方法;涉及较少特征的混合方法的推荐性能要优于涉及较多特征的混合方法。

关键词: 标签推荐; 统计语言模型; 隐含话题模型; 不同粒度
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)02-0504-06
doi:10.3969/j.issn.1001-3695.2012.02.028

Comparative research on different grain-based tag recommendation algorithm

JIN Yan-an[1,2], LI Yu-hua[1], LIU Xing-jun[2]
(1. *School of Computer Science, Huazhong University of Science & Technology, Wuhan* 430074, *China*; 2. *School of Information Management, Hubei University of Economics, Wuhan* 430205, *China*)

Abstract: Social tagging system has a characteristic that different entities from different grain have different descriptive power. This paper proposed some methods to recommend more precise tags from fine word-grained and coarse topic-grained according to this characteristic. The descriptions and tags of documents were modeled with statistic language model (fine word-grained) and latent dirichlet allocate model (coarse topic-grained), respectively. The paper hybrided different single model to recommend tags after using a single model, and then compared their different performances. The results of experiments show that the performance of word-grained tag recommendation is better than the topic-grained one, and the hybrid methods are better than non-hybrid ones, and the less the related features of hybrid are, the better the performance is obtained.

Key words: tag recommendation; statistic language model; latent dirichlet allocate model; different grain

0 引言

Web 2.0 下,社会标签系统作为信息资源组织的一种方式越来越受到网络用户的欢迎,但社会标签与生俱来的随意性和稀疏性削弱了标签在信息资源组织、检索、分享和利用的作用。因此,社会标签推荐的研究逐渐引起了学术界和企业界的广泛关注。标签推荐就是通过考察、分析、挖掘信息资源的内容和用户的标注历史、显式或隐式的关系为待标注信息资源提供一系列高质量的标签作为候选。推荐的目的主要是:a)简化标注程序,方便用户使用,从而增加社会标签系统的可用性和粘性;b)提高标签的质量,降低错拼、歧义等情况,提高标签在信息资源组织、检索、利用和发现的作用;c)改变标签空间的结构,使得标签空间更快地稳定和收敛,进而涌现语义。

文献[1]通过研究发现用户标注信息资源的首要动机是标志信息资源关于什么内容的或者是什么。因此,在对社会标签进行推荐时,应该考虑的第一个方面就是要分析信息资源的内容。这样就涉及到信息资源的内容如何表示的问题。对于文本形式的信息资源(如 Web 页面)来说,很自然地把文本中出现的词作为表示信息资源最细粒度的特征,把承载内容的语义信息(隐含话题)作为表示信息资源较粗粒度的特征。在细粒度上,词可以是信息资源的标题、描述、正文内容中的词或者社会标签。对于非文本形式的信息资源,因为不存在词这样细粒度的特征,所以只能对其上下文或者已有社会标签的集合进行分析,将它们所包含的词作为最细粒度的特征,而同样将承载内容的语义信息(隐含话题)视为较粗粒度的特征。

众多用户的标签汇聚在一起时,就会涌现出语义,不同用户选择相同的词条作为社会标签时,可能代表着不同的含义。在进行推荐时,如果对这些标签从语义上不加以区分,就会造成推荐的结果不令人满意。另一方面,词作为最常用的文档特征时常存在由于粒度太细而造成的稀疏性,如"苍山""洱海""下关"这三个词条。实际上此信息资源蕴涵着"大理旅游"这一主题,如果使用"大理"这一词进行标注,就可能克服词的稀疏性。因此,针对具体的信息资源,进行不同粒度(词粒度和隐含的主题粒度)的标签推荐是一个值得研究的课题。

在进行细粒度推荐时,可以直接利用内容进行推荐。Lipczak 等人[2]使用论文标题、其他用户使用的标签和用户自

收稿日期: 2011-06-23; **修回日期**: 2011-08-16 **基金项目**: 国家自然科学基金资助项目(70771043);湖北省教育科学"十一五"规划项目(2010B039);湖北省人文社科项目(2010Q094)

作者简介: 靳延安(1975-),男,河南郑州人,讲师,博士,CCF 会员,主要研究方向为 Web 智能处理、数据挖掘等(yan.an.jin@hbue.edu.cn);李玉华(1968-),女,副教授,博士,主要研究方向为数据挖掘、机器学习;刘行军(1975-),男,讲师,硕士,主要研究方向为知识管理、数据挖掘.

己的标签为论文进行了标注,他们的工作在 RSDC Challenge 中获得了基于内容进行推荐的第一名。Heymann 等人[3]使用文档的正文和内链接信息为文档预测了标签。词袋模型[4]也经常被用来表示文档特征,采用该表示方法的学者往往把标签预测的问题视为文本分类问题[5],每一个标签就是一个分类。Ohkura 等人[6]为每个标签训练了一个文本分类器来判定每一个文档是否适用该标签,从而解决了由于文档标签稀疏性而带来的不能辅助用户浏览文档的问题。一般来讲,由于文档中标签的分布是符合幂率分布的,这种分布将严重影响训练样本的不平衡性。因此,基于文本分类的方法只适用于少量的标签,并且需要训练大量的分类器。Sigma 和 Andy 利用 Web 2.0 的集体智慧和词的语义特征通过关键字提取、语义处理和人工神经网络学习三个步骤来推荐标签[7]。Shepitsen 等人[8]在层次聚类的基础上提出了一个个性化标签推荐框架,该框架不依赖于具体的聚类算法,但依赖于用户浏览的上下文环境。实际上,也可以间接使用内容进行推荐。Steffen 等人[9]使用 PITF 模型、使用其他描述集和标签集来建模文档,获得了更好的推荐性能。

为了克服细粒度词特征稀疏的问题,粗粒度的隐含主题模型[10,11]也被用于标签推荐的相关工作。Landauer 等人在文献[11]中详细介绍了最早的隐含主题模型—隐含语义分析方法(latent semantic analysis, LSA),其主要思想就是对传统的 TFIDF 方法进行扩充,引入了 SVD、向量化词和文档,进而方便词和文档的比较。之后,由 Hofmann 改进为基于概率的 PLSA 模型[10]。目前,标签推荐中使用最为广泛的模型是 Blei 等人[12]提出的隐含主题模型(latent dirichlet allocate, LDA)。LDA 模型是一种无监督的概率图模型,它将文档表示为 K 个隐含主题上的一个分布,而文档中的每个词都是由一个不可观察的隐含主题生成,这些隐含主题则是从文档对应的分布中采样得到。LDA 相对于 LSA 和 PLSA 来说有着更好的理论基础,其每个参数分布都有对应的先验分布,这些先验分布使得 LDA 能够据此推断未见文档的主题分布。LDA 等模型建模的对象是文档中的词。如果将文档中的标签也加入到模型中,则隐含主题模型也可以用于标签推荐。Blei 等人[13]在 LDA 的基础上增加了一个连续的变量代表外部的标注,并在训练时利用有标注的数据来得到最优的参数,这就是有监督的隐含主题模型(Supervised LDA)。在标签推荐方面,许多学者直接应用 LDA[14,15]或者引进 LDA 参数进行了研究[16,17]。

本文以信息资源的描述信息和标签为研究对象,使用基于词特征的统计语言模型(language model,LM)和基于语义特征的隐含话题模型(LDA)、标签—话题模型(tag-topic,TT)、用户—标签—话题模型(user-tag-topic,UTT)分别建模信息资源的描述信息和标签并推荐标签。最后,比较考查了建模方法的互补性、优劣性和适用场合。

1　基于词粒度的标签推荐算法

1.1　统计语言模型

统计语言模型广泛应用在自然语言处理领域中,如语音识别、机器翻译和信息检索。本质上讲它,是一个概率分布模型,主要描述自然语言的统计和结构方面的内在规律。假设有一个符号的集合 W,每一个$\{w_i\}$称做一个“单词”,由零个或多个单词连接起来就组成了一个字符串 $S=w_1w_2,\cdots,w_n$,字符串可长可短,如实际语言中的句子、段落或者文档都可以看做一个字符串。所有合法的字符串的集合称做一个语言,一个语言模型就是这个语言里的字符串的概率分布模型。

在信息检索中,统计语言模型最基本的思想就是把一个查询 q 和文档 d 的相关性解释为从文档中产生一个查询的概率模型[18],如式(1)所示。

$$P_{\mathrm{LM}(q|d)}=\prod_{w\in q}p(w|d) \tag{1}$$

其中:w 是查询中的一个词。$p(w|d)$是从文档 d 中产生查询词 w 的概率,这个概率可以按式(2)进行计算:

$$p(w|d)=\frac{N_d}{N_d+\lambda}\times\frac{tf(w,d)}{N_d}+\left(1-\frac{N_d}{N_d+\lambda}\right)\times\frac{tf(w,D)}{N_D} \tag{2}$$

其中:N_d 为文档 d 以词为单位计算出的长度;$tf(w,d)$是文档 d 中词 w 出现的词频;N_D 是所有的文档集中词的总数;$tf(w,D)$是所有的文档集中词 w 出现的词频;λ 是一个狄雷克雷特平滑因子,它的取值根据文档集中的平均文档长度来设定即 $\lambda=N_d/N_D$。

1.2　词粒度标签推荐基本思想

词粒度标签推荐的本质是把信息资源的内容看做是由较细粒度的词组成,推荐就是根据词的分布将信息资源归入相应的分类即所推荐标签。换句话说,就是给定一个信息资源的内容,它产生一个标签的概率是多少? 这个概率类似于统计语言模型中从文档 d 中产生查询词 w 的概率,因此,可以依式(2)计算。

网络信息资源可能是 Web 页面,也可能是图片或视频材料,为了不失研究的一般性,本文采用信息资源的描述集 D 和标签集 T 作为不同形式的内容来研究词粒度的标签推荐。根据不同形式的内容提出四种不同的模型来推荐标签,即基于描述集的统计语言模型(word language model,wLM)、基于标签集的统计语言模型(tag langugage model,tLM)、描述集和标签集的简单混合统计语言模型(simple combination word with tag language model,SCLM)以及描述集和标签集的混合相关模型(combination word with tag language related model,CRM)。

1.3　基于描述集的统计语言模型

对于一个网络信息资源 r, 用户 u 会使用一个描述集 w_u 来描述它,描述集 w_u 可能包括 1 个以上的描述词 w_{ui}。为了方便处理,把不同用户对同一信息资源 r 的描述集 w_u 进行合并,合并时允许描述词 w_{ui} 存在重复,但要去除非英文字符、停止词、数字和标点符号,这样就组成一个标签推荐的候选集 D_r。因此,每一个网络信息资源可以形式化表示为

$$D_r=\bigcup_i w_u=\bigcup_{1\leqslant u\leqslant n,0\leqslant j\leqslant m}t_{uj}$$

其中:w_{ij}是用户 i 描述资源 r 所使用的第 j 个描述词;n 是用户数量;m 为用户 i 对资源 r 描述集中描述词的个数。

根据式(2),资源 r 的描述集 D_r 产生标签 t 的概率可以用式(3)来计算:

$$P_{\mathrm{wLM}}(t|r)=\frac{N_{D_r}}{N_{D_r}+\lambda}\times\frac{tf(t,D_r)}{N_{D_r}}+\left(1-\frac{N_{D_r}}{N_{D_r}+\lambda}\right)\times\frac{tf(t,D)}{N_D} \tag{3}$$

其中:D 为所有信息资源的描述集的集合;N_{D_r}是信息资源 r 的描述集中词的个数;N_D 是所有的信息资源描述集中词的个数;λ 是狄雷克雷特平滑因子;$tf(t,D_r)$是信息资源 r 的描述集中出现标签 t 的次数;$tf(t,D)$是所有信息资源的描述集中出现标

签 t 的次数。

1.4　基于标签集的统计语言模型

网络信息资源除了可以用其正文表示其内容外，还可以用标注时所赋予的标签来表示，即将所有用户的标签合并起来作为网络信息资源的表示。

对于网络信息资源 r，用户 u 都会使用一些标签 t_u 来标注信息资源 r，标签集 t_u 可能包括 1 个以上的标签 t_{ui}。把不同用户对同一信息资源 r 的标签集 t_u 进行合并，组成一个标签推荐的候选集 T_r。因此，每一个信息资源可以形式化表示为

$$T_r = \bigcup_i t_u = \bigcup_{1\leqslant u\leqslant n,0\leqslant j\leqslant m} t_{uj} \tag{4}$$

其中：t_{uj} 是用户 u 标注资源 r 所使用的第 j 个标签；n 是用户数量；m 为用户 u 对资源 r 标注时使用的标签个数。

同样，根据原始的统计语言模型，信息资源 r 的标签集 T_r 产生标签 t 的概率可以用式(5)来计算。

$$P_{\text{tLM}}(t|r) = \frac{N_{T_r}}{N_{T_r}+\lambda}\times\frac{tf(t,T_r)}{N_{T_r}} + \left(1-\frac{N_{T_r}}{N_{T_r}+\lambda}\right)\times\frac{tf(t,T)}{N_T} \tag{5}$$

其中：T 为所有信息资源的描述集的集合；N_{T_r} 是信息资源 r 的描述集中词的个数；N_T 是所有的信息资源的描述集中词的个数；λ 是狄雷克雷特平滑因子；$tf(t,T_r)$ 是信息资源 r 的描述集中出现标签 t 的次数；$tf(t,T)$ 是所有信息资源的描述集中出现标签 t 的次数。

1.5　描述集和标签集的简单混合统计语言模型

实际上，描述集和标签集可以同时用来表示信息资源的内容。描述集可以认为是信息资源的摘要，标签集可以看做信息资源的语义标注信息，故两者均可作为信息资源推荐标签的候选集。在这种情况下，如果一个词在描述集中占的权重比较高，在标签集中占的权重也比较高，那么这个词被推荐给信息资源的可能性就非常大；如果一个词在标签集中占的权重特别小，那么势必会削弱该词在描述集中所占的地位，相反亦然。因此，可以使用一种简单的方法聚合基于描述集推荐和基于标签集推荐的模型，聚合如式(6)所示。

$$P_{\text{SCLM}}(t|r) = P_{\text{wLM}}(t|r)P_{\text{tLM}}(t|r) \tag{6}$$

1.6　描述集和标签集的混合相关模型

在 SCLM 模型中，存在这样的可能：如果一个词在描述集中所占的比重非常大，但是在标签集中所占的比重为零，这样聚合的结果是零。显然，该词是不可能作为标签被推荐。直接从 SCLM 的聚合公式也可以看出，该模型要求词在描述集和标签集中都有一定的权重，这不符合现实情况。实际上，文献[12]通过实验证明，标签是最能体现用户的意图和信息资源的内容。因此，推荐的标签还是应该从信息资源以前的标签集中产生。另外，如果一个标签和描述集中重要的描述词有更强的相关性，那么推荐这个标签的可能性就更高。基于这些假设，采用如式(7)推荐标签：

$$P_{\text{CRM}}(t|r) = \sum_{w\in D_r,t\in T_r} p(w)s(w,t) \tag{7}$$

其中：$p(w)$ 是描述词 w 在描述集中的权重，而 $s(w,t)$ 则是描述词 w 与标签 t 之间的相关性。这样就将问题转换成了如何更准确地计算描述词在描述集中的权重 $p(w)$ 及描述词和标签的相关性 $s(w,t)$。

描述词 w 在描述集中的权重 $p(w)$ 可以通过信息检索领域中广泛应用的 TFIDF 来计算，如式(8)所示。

$$P(w) = \log\left(\frac{tf(w,D_r)}{N_{D_r}}+1\right)\log\left(\frac{N_D}{|D_{r'}|w\in D_{r'}|}+1\right) \tag{8}$$

描述词和标签的相关性计算有许多方法，最简单直观的是计算描述词和标签共同出现的次数占所有情况的比例。但是，由于标签的稀疏性，直接使用该方法可能使得相似性无法得到计算。因此，使用改进的 Google 距离公式[19]计算描述词和标签的相关性。

$$s(w,t) = \frac{\max(\{\log f(w),\log f(t)\}) - \log f(w,t)}{\log N - \min(\{\log f(w),\log f(t)\})} \tag{9}$$

其中：$f(w)$、$f(t)$ 分别为描述词和标签在描述集和标签集的并集中出现的次数；$f(w,t)$ 为描述词和标签同时出现在并集中的次数；N 为并集总计词数。

2　基于话题粒度的标签推荐算法

2.1　LDA 模型

LDA 模型是一个广泛应用于文本话题检测的三层贝叶斯网络生成模型。生成模型是指该模型可以随机生成可观测的数据，LDA 可以随机生成一篇由 N 个主题组成的文章。通过对文本的建模，可以对文本进行主题分类，判断相似度等。LDA 中通过将文本映射到主题空间，即认为一篇文章由若干主题随机组成，从而获得文本间的关系。LDA 模型认为文档是忽略了语法和出现顺序关系的词的集合。

LDA 的建模过程是逆向通过文本集合建立生成模型。例如：假设一个语料库中有三个主题，即体育、科技、电影。一篇描述电影制作过程的文档，可能同时包含主题科技和主题电影，而主题科技中有一系列的词，这些词与科技有关，并且它们有一个概率，代表的是在主题为科技的文章中该词出现的概率。同理，在主题电影中也有一系列与电影有关的词，并对应一个出现概率。当生成一篇关于电影制作的文档时，首先随机选择某一主题，选择到科技和电影两主题的概率更高；然后选择单词，选择到那些与主题相关的词的概率更高，这样就完成了一个单词的选择。不断选择 N 个单词，这样就组成了一篇文档。

具体来说，生成一篇文档按照如下步骤[12]：

a)选择 N，N 服从 Poisson(ξ)分布，N 代表文档的长度。

b)选择 θ，θ 服从 Dirichlet(α)分布，代表的是某个主题发生的概率，α 是 Dirichlet 分布的参数。

c)对 N 个单词中的每一个：

(a)选择主题 z_n，z_n 服从 Multinomial(θ)多项分布。z_n 代表当前选择的主题。

(b)选择 w_n，根据 $p(w_n|z_n,\beta)$：在 z_n 条件下的多项分布。

上式中，β 是一个 $K\times V$ 的矩阵，K 是 Dirichlet 分布的维度即话题的维度，$\beta_{ij} = P(w_j = 1|z_i = 1)$，也就是说 β 记录了某个主题条件下生成某个单词的概率。它代表的概率模型如式(10)所示。

$$p(\theta,z,w|\alpha,\beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n,\beta) \tag{10}$$

2.2　隐含话题粒度标签推荐基本思想

可以按如下过程来理解隐含话题模型：当一个人要生成一个文档时，他的脑海中肯定有一些主题，为了阐述某个主题，他就需要从这个主题所包含的所有词中以一个可能的概率选择一个表示主题的词。这时，整个文档就表示为不同主题的混合。当文档的作者只有一个人时，这些主题就反映了作者的观

点和他特有的词典。在社会标签系统中,允许多个用户同时标注信息资源,多个用户所产生的话题反映了大家对信息资源所持有的共同的观点,所使用的标签反映了大家在描述信息资源时共用的词典。

LDA 建模过程可以看做是发现每个信息资源的混合主题 $p(z|r)$,每个主题可以看做是符合词的混合分布 $p(t|z)$。这一过程可以形式化为[12]

$$p(t_i|r_i)=\sum_{j=1}^{z}p(t_i|z_i=j)p(z_i=j|r_i) \tag{11}$$

其中:$p(t_i|r_i)$是信息资源 r_i 中出现词 t_i 的概率;z_i 是隐含话题;$p(t_i|z_i=j)$是隐含话题 j 中词 t_i 出现的概率;$p(z_i=j|r_i)$为信息资源 r_i 时关于话题 j 的概率。z 是预先定义的隐含话题的数目,该值用于调整隐含话题的具体化程度。LDA 的任务是从没有标注的信息资源中使用狄雷克雷特先验估计话题—词分布 $p(t|z)$和信息资源—话题分布。LDA 使用 Gibbs 采样来达到这个目的,即 Gibbs 为信息资源中的每个词进行多次迭代,每次迭代都以式(12)为词 t_i 采样一个新的话题,直到 LDA 模型的参数收敛。

$$p(z_i=j|t_i,r_i,z_{-i})\propto\frac{C_{t_{ij}}^{TZ}+\beta}{\sum_{t\in T}C_{t_{ij}}^{TZ}+|T|\beta}\times\frac{C_{r_{ij}}^{RZ}+\alpha}{\sum_{z\in\mathbf{Z}}C_{r_{ij}}^{TZ}+|z|\alpha} \tag{12}$$

其中:C^{TZ}、C^{RZ}记录了所有的话题—词和资源—话题分配的数量;z_{-i}表示除了当前话题的所有话题—词和资源—话题分配;α、β是狄雷克雷特先验的超参数,是用来平滑模型的。这样,式(11)的后验概率可以按式(13)(14)计算:

$$p(t_i|z_i=j)=\frac{C_{t_{ij}}^{TZ}+\beta}{\sum_{t\in T}C_{t_{ij}}^{TZ}+|T|\beta} \tag{13}$$

$$p(z_i=j|r_i)=\frac{C_{r_{ij}}^{RZ}+\alpha}{\sum_{z\in Z}C_{r_{ij}}^{RZ}+|Z|\alpha} \tag{14}$$

根据式(11),可以得到一个给定的信息资源产生一个标签的概率为

$$p(t_i|r_i)=\sum_{j=1}^{z}p(t_i|z_i=j)p(z_i=j|r_i)=\frac{C_{t_{ij}}^{TZ}+\beta}{\sum_{t\in T}C_{t_{ij}}^{TZ}+|T|\beta}\times\frac{C_{r_{ij}}^{RZ}+\alpha}{\sum_{z\in Z}C_{r_{ij}}^{RZ}+|Z|\alpha} \tag{15}$$

2.3　基于描述集的 LDA 模型

一般来讲,LDA 模型建模的对象是文档的正文,隐含话题建立起了文档和词的联系。但是,正如前面所述及的原因,本文使用信息资源的描述集来代替信息资源的正文。这样,文档和词的联系就转换为描述集和描述词的关系。首先按照 1.4 节中的方法处理描述集,对得到的每个信息资源的描述集,使用式(16)来计算产生某个标签的概率。

$$p_{\mathrm{wLDA}}(t_i|D_r)=\sum_{j=1}^{z}p(t_i|z_i=j)p(z_i=j|D_r)=\sum_{j=1}^{z}\frac{C_{t_{ij}}^{D_rZ}+\beta}{\sum_{t\in D_r}C_{t_{ij}}^{TZ}+|T|\beta}\times\frac{C_{D_rj}^{DZ}+\alpha}{\sum_{z\in Z}C_{D_rz}^{DZ}+|Z|\alpha} \tag{16}$$

2.4　基于标签集的 LDA 模型

为了验证 LDA 模型的效果,还可以使用标签集作为信息资源表示的另外一种形式,这样可以使用隐含话题将信息资源和标签之间建立联系。对标签集按照 1.5 节中的处理方法,得到每个信息资源的标签集,使用式(17)来计算产生某个标签的概率。

$$p_{\mathrm{tLDA}}(t_i|T_r)=\sum_{j=1}^{z}p(t_i|z_i=j)p(z_i=j|T_r)=\sum_{j=1}^{z}\frac{C_{t_{ij}}^{TZ}+\beta}{\sum_{t\in T}C_{t_{ij}}^{TZ}+|T|\beta}\times\frac{C_{T_rj}^{TZ}+\alpha}{\sum_{z\in Z}C_{T_rz}^{TZ}+|Z|\alpha} \tag{17}$$

2.5　标签—话题模型

LDA 是一个概率隐含变量模型,它把信息资源作为主题的混合,信息资源词的选择是根据信息资源的主题来选择。产生过程中文档—话题和话题—词的分布均服从狄雷克雷特先验,所以信息资源之间相互独立。

为了给信息资源推荐最合适的标签,必须弄清楚标签和信息资源的联系。在社会标签系统中,多个用户可以标注同一个信息资源,可以认为标签是和描述词同样地产生关系。标准的 LDA 模型只建模信息资源、主题和词。为了把标签引入标准的 LDA 模型,同时建模信息资源的描述集和标签集,本节在文献[20]的基础上改进 Author-Topic 模型,提出了新的模型 Tag Topic(TT)来进行标签的推荐,其本质是将描述集和标签集进行合并处理。

产生标签的概率可以用式(18)计算:

$$p_{\mathrm{TT}}(t_i|T_r)=\sum_{j=1}^{z}p(t_i|z_i=j)p(z_i=j|(T_r\cup D_r))=\sum_{j=1}^{z}\frac{C_{t_{ij}}^{(T_r\cup D_r)Z}+\min(\beta,\beta_t)}{\sum_{t\in(T_r\cup D_r)}C_{t_{ij}}^{(T_r\cup D_r)Z}+|T_r\cup D_r|\min(\beta,\beta_t)}\times\frac{C_{T_rj}^{TZ}+\alpha}{\sum_{z\in Z}C_{T_rz}^{TZ}+|Z|\alpha} \tag{18}$$

2.6　用户—标签—话题模型

在 TT 模型中,同时建模了描述集和标签集,但这并不符合实际情况,因为社会标签系统中还存在着用户的角色,必须同时建模描述集、标签集和用户。于是,在文献[21]的基础上改进 Author-Conference-Topic 模型并提出 User-Tag-Topic(UTT)模型。该模型产生标签的基本过程:在符合均匀分布 γ 的用户中选择一个特定用户 u,用户 u 从自己的符合多项分布的话题列表中选择一个话题 z,接下来从符合多项分布的词表中选择词 w。产生标签的概率可以用式(19)计算:

$$p_{\mathrm{UTT}}(t_i|T_r)=\sum_{j=1}^{z}\sum_{k=1}^{u}p(t_i|z_i=j)p(z_i=j|u_k=j)p(u_k=j|T_r)=\sum_{j=1}^{z}\frac{C_{t_{ij}}^{(T_r\cup D_r)Z}+\min(\beta,\beta_t)}{\sum_{t\in(T_r\cup D_r)}C_{t_{ij}}^{(T_r\cup D_r)Z}+|T_r\cup D_r|\min(\beta,\beta_t)}\times\frac{C_{U_rz}^{UZ}+\alpha}{\sum_{z\in Z}C_{U_rz}^{UZ}+|U|\alpha}\times\frac{C_{T_ru}^{TU}+\alpha}{\sum_{u\in U}C_{T_ru}^{TU}+|U|\alpha} \tag{19}$$

2.7　基于混合粒度的标签推荐

在词粒度上进行推荐,关注的是信息资源正文、描述集和标签集中词本身,这个粒度的方法可以称为基于关键字的方法(keyword based approach, KBA)。该方法使用统计语言模型来建模,统计语言模型只关心具体的词而有不能建模话题的限制。隐含话题粒度的推荐更关注信息资源的概念知识或者说是语义知识,可以称这个粒度的方法为基于语义的方法(semantic based approach, SBA)。基于语义的隐含话题模型消除建模话题的限制,并可在多个因素上进行建模。为了充分发挥以上两种粒度推荐方法的各自优势,本文融合两种粒度进行推荐。直觉上讲,会提高推荐的性能。本节提出使用一个聚合函数来混合基于 KBA 和 SBA 的标签推荐方法。聚合函数如式(20)所示。

$$\eta(r,t)=\zeta(\delta(r,t),\vartheta(r,t)) \tag{20}$$

其中:$0\leqslant\delta(r,t)\leqslant1$ 是资源 r 基于 KBA 的标签推荐方法产生标签 t 的概率;$0\leqslant\vartheta(r,t)\leqslant1$ 是资源 r 基于 SBA 的标签推荐方法产生标签 t 的概率;$\zeta(,)$是聚合 KBA 和 SBA 的函数,该函数

符合以下公理：

公理 1 对于一个标签 t 和一个资源 r，如果 $\delta(r,t)=0$，则 $\eta(r,t)=\vartheta(r,t)$；如果 $\vartheta(r,t)=0$，则 $\eta(r,t)=\delta(r,t)$。

公理 2 对于一个标签 t 和一个资源 r，$\zeta(,)$ 应能保证 $0\leqslant\eta(r,t)\leqslant1$。

公理 3 对于一个资源 r 和两个标签 t_1、t_2，如果 $\delta(r,t_1)=\delta(r,t_2)$ 且 $\vartheta(r,t_1)\geqslant\vartheta(r,t_2)$，则 $\eta(r,t_1)\geqslant\eta(r,t_2)$；如果 $\delta(r,t_1)\geqslant\delta(r,t_2)$ 且 $\vartheta(r,t_1)=\vartheta(r,t_2)$，则 $\eta(r,t_1)\geqslant\eta(r,t_2)$。

公理 1 规定了聚合函数的边界，当 $\delta(r,t)=0$，意味着标签 t 在训练集中但不在基于 KBA 的候选集中，这时推荐的效果就等于是使用 SBA 的方法，所以 $\eta(r,t)=\vartheta(r,t)$。当 $\vartheta(r,t)=0$，意味着标签 t 不在训练集中但在基于 KBA 的候选集中，这时推荐的效果就等于是使用基于 KBA 方法，所以 $\eta(r,t)=\delta(r,t)$。公理 2 保证概率计算结果是在[0,1]。公理 3 说明了基于 KBA 和 SBA 方法的相互作用。实际上，聚合函数与应用无关，如何选择聚合函数与本文无关，这里不作说明。本文采用式(21)来聚合 KBA 和 SBA 方法。

$$\eta(r,t)=\xi\delta(r,t)+(1-\xi)\vartheta(r,t) \tag{21}$$

基于以上讨论，对基于 KBA 和 SBA 的方法进行混合推荐。在 KBA 方法中使用 LM 模型，在 SBA 方法中使用 LDA、TT、UTT。本文主要考虑以下聚合：wLM + wLDA(式(22))、wLM + tLDA(式(23))、wLM + TT(式(24))、wLM + UTT(式(25))、tLM + wLDA(式(26))、tLM + tLDA(式(27))、tLM + TT(式(28))和 tLM + UTT(式(29))。

$$p(t|r)=\xi p_{\mathrm{wLM}}(t|r)+(1-\xi)p_{\mathrm{wLDA}}(t|D_r) \tag{22}$$

$$p(t|r)=\xi p_{\mathrm{wLM}}(t|r)+(1-\xi)p_{\mathrm{tLDA}}(t|T_r) \tag{23}$$

$$p(t|r)=\xi p_{\mathrm{wLM}}(t|r)+(1-\xi)p_{\mathrm{TT}}(t|T_r) \tag{24}$$

$$p(t|r)=\xi p_{\mathrm{wLM}}(t|r)+(1-\xi)p_{\mathrm{UTT}}(t|T_r) \tag{25}$$

$$p(t|r)=\xi p_{\mathrm{tLM}}(t|r)+(1-\xi)p_{\mathrm{wLDA}}(t|D_r) \tag{26}$$

$$p(t|r)=\xi p_{\mathrm{tLM}}(t|r)+(1-\xi)p_{\mathrm{tLDA}}(t|T_r) \tag{27}$$

$$p(t|r)=\xi p_{\mathrm{tLM}}(t|r)+(1-\xi)p_{\mathrm{TT}}(t|T_r) \tag{28}$$

$$p(t|r)=\xi p_{\mathrm{tLM}}(t|r)+(1-\xi)p_{\mathrm{UTT}}(t|T_r) \tag{29}$$

3 实验分析

3.1 实验设置

1)数据集

实验所用数据集来自于 2008 年 ECML/PKDD Discovery Challenge 竞赛所用的 Bibsonomy 数据集。原始数据集中共有 33 256 个字词，13 276 个标签 tags，1 185 个用户，14 443 个文档和 262 445 个三元组。去掉原始数据集非英语词、停止词并进行词干化处理后，得到 18 450 个字词、10 702 个标签、861 个用户、13 179 个文档和 181 491 个标注。

2)对照方法

实验主要目的是发现不同的内容(描述集、标签集和用户集)、不同的粒度(词粒度和话题粒度)对推荐性能的影响，所以主要比较 KBA 所包含的 wLM、tLM、SCLM、CRM 和 SBA 所包含的 wLDA、tLDA、TT、UTT 以及它们的混合方法 wLM + wLDA、wLM + tLDA、wLM + TT、wLM + UTT、tLM + wLDA、tLM + tLDA、tLM + TT 和 tLM + UTT。

3)参数设置

本文对基于词粒度的推荐方法和基于语义的方法进行了大量的比较研究，每一种方法都有大量的参数进行设置。对实验中所涉及到的参数设置如下：

基于词粒度的推荐方法主要采用统计语言模型。统计语言模型中的狄雷克雷特平滑因子 λ 是非常重要的参数，对推荐性能有较大影响，实验中一般设为文档的平均长度。对于描述集，λ 设为 15；对于标签集，λ 设为 5。基于语义的方法主要采用隐含话题模型 LDA、标签—话题模型 TT、用户—标签—话题模型 UTT。话题模型中很重要是超参数的估计。超参数的估计一般使用期望最大化方法 EM 和吉布斯采样 Gibbs 方法实现。但是由于期望最大化方法有收敛于局部最大化和计算效率低的特点，本文采样吉布斯采样方法实现。实验中，设置超参数 α、β、λ 分别为 50/|T|、0.01 和 0.1。根据话题模型的指标 perplexity[12] 来测量与比较，设置话题数目为 20200 进行对比。对于混合方法，设置 ξ=0.15。

4)评价方法

一般来讲，推荐算法的评价方法可以分为两类：a)推荐的精准性，即比较推荐值和实际值之间的差异，通常会使用均绝误差(mean absolute error，MAE)和均方根误差(root mean square error，RMSE)；b)排序的精准性，即产生与用户排序一致性能力的度量，通常包括 Top k 推荐[22] 和 NDCG[23]。本文主要目的是向用户推荐前 k 个标签，因此，采用文献[22]中的评价方法来评价，具体方法见文献[22]。

3.2 推荐 Top k 标签的性能比较

表 1 给出了本文所提出的混合模型和对照方法模型推荐标签的性能，包括词粒度方法、语义粒度方法和两者的混合方法。从表中可以看出：就单一方法讲，基于词粒度模型的推荐性能要比语义的模型的推荐性能好，如 wLM、tLM、SCLM、CRM 要好于 wLDA、tLDA、TT、UTT；基于两种方法的混合方法的性能要好于没有混合的基于话题的单个方法，如 wLM + wLDA、wLM + tLDA、wLM + TT、wLM + UTT、tLM + wLDA、tLM + tLDA、tLM + TT 和 tLM + UTT 要好于 wLDA、tLDA、TT、UTT。但也有例外，比如推荐性能最好的词粒度方法的 wLM(70%)要高出混合方法中性能最好的 wLM + tLDA(65%) 5%；同时，可以看出，涉及较少特征的混合方法的推荐性能要优于涉及较多特征的混合方法，比如 wLDA(55%)和 tLDA (57%)要优于 TT(52%)和 UTT(47%)；wLM + wLDA(64%)和 tLM + tLDA(65%)的推荐性能要优于 tLM + TT(63%)和 tLM + UTT(57%)。

表 1 不同推荐方法推荐标签的性能比较

方法	话题数目					平均精确度
	2	4	6	8	10	
wLM	.8249	.7649	.6750	.6336	.6022	.7001
tLM	.8550	.7373	.6684	.5773	.6000	.6876
wLM + tLDA	.7915	.7013	.6112	.6064	.6253	.6671
tLM + tLDA	.8746	.5661	.6697	.6161	.5676	.6588
wLM + wLDA	.8648	.5687	.6698	.6127	.5675	.6567
wLM + TT	.7068	.6338	.6252	.6552	.6543	.6551
tLM + TT	.8588	.7317	.6578	.4472	.5157	.6422
CRM	.7693	.6126	.6220	.6695	.4806	.6308
tLM + wLDA	.7622	.6210	.6158	.6688	.4822	.6300
SCLM	.7623	.6298	.6187	.5845	.5269	.6244
wLM + UTT	.8074	.7160	.6380	.4335	.5032	.6196
wLDA	.7400	.6575	.5449	.5563	.5379	.6073
tLDA	.7923	.4158	.6103	.5590	.5195	.5794
tLM + UTT	.7068	.6338	.6252	.5582	.4601	.5568
TT	.5031	.4833	.5587	.6289	.4374	.5223
UTT	.5102	.4654	.4878	.5367	.3692	.4739

出现如上所述现象的主要原因是：a)语义粒度(话题模

型)隐去了词粒度的很多细节,这样就造成基于词粒度模型的推荐性能要比语义的模型的推荐性能好;b)混合的方法由于增加了资源的描述维度,在一定意义上使资源的多方面主题得以体现,所以基于两种方法的混合方法的性能要好于没有混合的基于话题的单个方法;c)如果是在建模时,引入太多的特征,可能会给建模带来更多的噪声,使得推荐的性能下降,所以涉及较少特征的混合方法的推荐性能要优于涉及较多特征的混合方法。

4　结束语

标签推荐就是对一个新的文档,根据社会标签的历史数据,由算法找出最适合的标签。从推荐的依据来说,标签推荐问题可以分为与内容无关的和基于内容的两类。与内容无关的标签推荐不考虑文档的内容,只依靠其他用户为相同文档标注的标签,或是本用户曾经使用过的标签。与内容无关的标签推荐由于无须分析文档内容,多用于照片、音乐等无法进行有效内容理解的多媒体文档。但是,与内容无关的方法不能给无人标注过的新文档推荐标签,对于关注较少的文档推荐效果也不好。关注和标注较多的文档往往是整个文档集合中很小的一部分,所以与内容无关的标签推荐方法虽然效果较好,但是应用范围受到了很大的限制。

基于内容的标签推荐使用文档的内容作为推荐标签的依据,一般来说使用的是文本内容。基于内容的方法不受新文档和冷门话题的限制,对于有充足文本内容的文档来说更适用,如网页、博客文章、新闻等。本文针对不同粒度特征进行推荐的问题,使用统计语言模型和隐含话题模型分别对描述集、标签集进行建模,提出融合不同的特征粒度来进行推荐。首先建立了混合不同粒度方法应该遵循的公理系统,然后对基于细粒度的词特征推荐和基于粗粒度的隐含话题推荐进行了两两聚合。实验结果表明:就单一方法讲,基于词粒度模型的推荐性能要比语义的模型的推荐性能好;基于两种方法的混合方法的性能要好于没有混合的基于话题的单个方法;涉及较少特征的混合方法的推荐性能要优于涉及较多特征的混合方法。

参考文献:

[1] Al-KHALIFA H S, DAVIS H C. Measuring the semantic value of folksonomies[C]//Proc of the 2nd International IEEE Conference on Innovations in Information Technology. [S. l.]: IEEE,2006:17-21.

[2] LIPCZAK M, HU Yu-ming, KOLLET Y, *et al*. Tag sources for recommendation in collaborative tagging systems[C]//Proc of ECML/PKDD Discovery Challenge. [S. l.]: CEUR Workshop Preceedings, 2009:157-172.

[3] HEYMANN P, RAMAGE D, GARCIA-MOLINA H. Social tag prediction[C]//Proc of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press,2008:531-538.

[4] LEWIS D D. Naive (Bayes) at Forty: the independence assumption in information retrieval[C]//Lecture Notes in Computer Science. London, UK: Springer-Verlag,1998: 4-15.

[5] TATU M, SRIKANTH M, D'SILVA T. Tag recommendations using bookmark content[C]//Proc of ECML/PKDD Discovery Challenge Workshop, Part of the European Conference on Machine Learning and Principles and Practice of Know Cedge Discovery in Databases. 2008: 96-107.

[6] OHKURA T, KIYOTA Y, NAKAGAWA H. Browsing system for weblog articles based on automated folksonomy[C]//Proc of Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics at WWW2006. 2006.

[7] LEE S O K, ANDY H W C. A Web 2.0 Tag recommendation algorithm using hybrid ANN semantic structures[J]. International Journal of Computers,2007,1(3):49-57.

[8] SHEPITSEN A, GEMMELL J, MOBASHER B, *et al*. Personalized recommendation in collaborative tagging systems using hierarchical clustering[C]//Proc of the 2008 ACM Conference on Recommender Systems. New York: ACM Press,2008: 259-266.

[9] STEFFEN R, LARS S. Pairwise interaction tensor factorization for personalized tag recommendation[C]//Proc of the 3rd ACM International Conference on Web Search and Data Mining. New York: ACM Press,2008:81-90.

[10] HOFMANN T. Probabilistic latent semantic indexing[C]//Proc of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM,1999:50-57.

[11] LANDAUER T, FOLTZ P, LAHAM D. An introduction to latent semantic analysis[J]. Discourse Processes,1998,25(2):259-284.

[12] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research,2003,3(1):993-1022.

[13] BLEI D M, McAULIFFE J D. Supervised topic models[J]. Advances in Neural Information Processing Systems, 2008, 20: 121-128.

[14] KRESTEL R, FANKHAUSER P, NEJDL W. Latent dirichlet allocation for tag recommendation[C]//Proc of the 3rd ACM Conference on Recommender Systems. New York: ACM Press,2009:61-68.

[15] KRESTEL R, FANKHAUSER P. Language models and topic models for personalizing tag recommendation[C]//Proc of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Washington DC: IEEE Computer Society, 2010:82-89.

[16] MORGAN H, IAN R, MARK J. Improving social bookmark search using personalised latent variable language models[C]//Proc of the 4th ACM International Conference on Web Search and Data Mining. New York: ACM Press,2011: 485-494.

[17] TSAI F S. A tag-topic model for blog mining[J]. Journal of Expert System and Application,2011,38(5):5330-5335.

[18] PONTE J M, CROFT W B. A language modeling approach to information retrieval[C]//Proc of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press,1998: 275-281.

[19] CILIBRASI R, VITANYI P M B. The google similarity distance[J]. IEEE Trans on Knowledge and Data Engineering,2007,19(3): 370-383.

[20] ROSEN-ZVI M, GRIFFITHS T L, STEYVERS M, *et al*. The author-topic model for authors and documents[C]//Proc of the 20th Conference on Uncertainty in Artificial Intelligence. Virginia: AUAI,2004: 487-494.

[21] TANG Jie, ZHANG Jing, YAO Li-min, *et al*. ArnetMiner: extraction and mining of academic social networks[C]//Proc of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press,2008:990-998.

[22] KOREN Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]//Proc of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press,2008: 426-434.

[23] JARVELIN K, KEKALAINEN J. Cumulated gain-based evaluation of IR techniques[J]. ACM Trans on Information Systems,2002,20(4):422-446.