软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(3):476−488 [doi: 10.3724/SP.J.1001.2012.04001]

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

融合关系与内容分析的社会标签推荐*

张 斌[1+], 张 引[1], 高克宁[2], 郭朋伟[1], 孙达明[1]

[1](东北大学 信息科学与工程学院,辽宁 沈阳 110004)

[2](东北大学 计算中心,辽宁 沈阳 110004)

Combining Relation and Content Analysis for Social Tagging Recommendation

ZHANG Bin[1+], ZHANG Yin[1], GAO Ke-Ning[2], GUO Peng-Wei[1], SUN Da-Ming[1]

[1](College of Information Science and Engineering, Northeastern University, Shenyang 110004, China)

[2](Computing Center, Northeastern University, Shenyang 110004, China)

+ Corresponding author: E-mail: zhangbin@ise.neu.edu.cn

Zhang B, Zhang Y, Gao KN, Guo PW, Sun DM. Combining relation and content analysis for social tagging recommendation. *Journal of Software*, 2012,23(3):476−488. http://www.jos.org.cn/1000-9825/4001.htm

Abstract: Tagging is one of the most important ways to categorize or indexing information at the age of Web 2.0. To handle the disadvantages of tagging systems such as inconsistentcies, redundancy and incompleteness, tag recommendation methods improve the quality of tags by providing candidate tags. In order to further improve the quality of tag recommendations, a tag recommendation method is proposed which bases on a combined analysis of the relations of objects in a tagging system and the content of resources. An LDA based generative tagging system model TSM/Forc that models object relation and resource content in a combined way is introduced, together with a probabilistic based tag recommendation method and a Gibbs sampling based model parameter estimation approach. Experiment results show that the proposed method could provide more accurate recommendations than the latest methods.

Key words: Web 2.0; social tagging; tag recommendation; combined approach; LDA (latent Dirichlet allocation)

摘 要: 标签是 Web 2.0 时代信息分类与索引的重要方式.为解决标签系统所面临的不一致性、冗余性以及完备性等问题,标签推荐通过提供备选标签的方法来提高标签的质量.为了进一步提升标签推荐的质量,提出了一种基于标签系统中对象间关系与资源内容融合分析的标签推荐方法,给出了基于 LDA(latent Dirichlet allocation)的融合表示对象间关系与资源内容的标签系统生成模型 TSM/Forc,提出了一种基于概率的标签推荐方法,并给出了基于吉布斯(Gibbs)抽样的参数估计方法.实验结果表明,该方法可以提供比当前主流与最新方法更加准确的推荐结果.

关键词: Web 2.0;社会标签;标签推荐;融合方法; LDA (latent Dirichlet allocation)

中图法分类号: TP391 **文献标识码**: A

Web 2.0 强调用户参与的特性,将信息发布的主导权从网站管理者转移到了网络用户身上.由于网络用户通

* 基金项目: 国家自然科学基金(61073062); 辽宁省自然科学基金(20102060); 中央高校基本科研业务费资助(N090604010); 沈阳市科学技术计划项目(F11-264-1-33)

收稿时间: 2010-08-28; 定稿时间: 2011-02-17

常没有耐心了解一个相对复杂的信息分类体系,传统的基于固定分类体系的信息分类方法也不再适应 Web 2.0 时代的信息发布方式.作为对传统分类方法的替代,人们使用标签系统以便更加容易地对 Web 2.0 上的信息进行分类或索引.使用标签系统,用户可以使用任意的词汇对信息进行分类.这种易用性使标签系统很容易被用户接受和使用,并使标签成为了 Web 2.0 时代一种重要的信息组织方式.

另一方面,标签系统这种不受控制的使用方法也带来了相应的问题.由于用户可以使用任意词汇分类信息,出于对信息和词汇的不同理解,不同的用户便不太可能使用完全一致的方法分类相同或者相似的信息.研究结果表明,基于标签的分类系统通常难以保证分类的一致性,并面临着冗余性、不完备性等问题[1].标签使用次数与资源所具有的标签数量所服从的齐普夫分布,使得标签系统中存在着大量仅仅被使用过数次的标签以及仅仅包含数个标签的资源[2].这种情况给基于标签的分类系统在信息检索等领域中的实际应用带来了问题.

解决标签系统所面临问题的核心在于提升标签的质量.针对这一问题,研究人员对标签推荐方法进行了大量的研究[2].标签推荐方法通过为用户在进行基于标签的分类过程中提供高质量的标签备选,或者在使用标签系统提供的分类信息用于信息检索等任务时,为信息附加了当前不具备的额外的高质量标签等方法,提升信息所带有的标签的质量.截至目前,研究人员已经提出了大量的标签推荐方法[3−5],并在来自开放式 Web 2.0 标签系统如 del.icio.us 等的数据集上获得了比较好的效果.

当前的标签推荐方法大多依靠对标签系统中对象间的关系进行分析,即通过分析标签系统中的标签、用户以及资源这三者之间的关系,获得推荐标签所需要的知识并应用于推荐任务中.这些方法的有效性依赖于用户、标签以及资源之间存在相对稠密的关联关系.然而,这些要求在现实应用中却不一定能够得到满足.例如在博客系统中,用户不可以像在 del.icio.us 这类的开放标签系统中对任意的信息进行标签分类,而只能对自己发布的信息附加标签.类似的问题也存在于基于标签的企业数据空间系统中.这些限制使得用户、标签以及资源之间的关联关系变得稀疏,限制了基于关系的标签推荐方法的性能.

值得注意的是,这种基于关系的标签推荐方法仅仅利用了对象之间的关系,并没有对资源自身的特征,即用户为资源附加特定标签的原因进行很好的发掘和利用,进而限制了基于关系的标签推荐方法的最终效果.据此可以认为,对资源自身的特征(如用户为资源附加特定标签的原因)进行分析,将能够进一步地获取推荐标签所需要的知识.将标签系统中对象之间的关系与资源本身的内容进行融合分析,将可以提升标签推荐系统的效果.

本文提出了基于关系与内容融合分析的 Web 2.0 信息标签推荐方法.通过扩展 Latent Dirichlet Allocation (LDA)[6]主题模型,本文将用户、标签、资源这 3 种类型的对象和组成资源内容的词汇,利用潜在的主题关联起来,利用一个模型完整地描述了对象关系与资源内容,避免了使用不同模型分别描述对象关系与资源内容后进行合并的方法,实现了对标签系统中的对象间关系与资源内容的融合分析.该模型完全生成模型的特性,使其可以有效地利用贝叶斯层次结构进行参数估计,很好地避免了模型的过度拟合问题,并具有很好的可扩展性.实验结果表明,所提出的方法可以提供比当前主流与最新的方法更加准确的标签推荐结果,在合理的时间与空间开销下,进一步地提升了标签推荐的质量.

1 基于关系与内容融合的标签系统模型 TSM/Forc

为了实现对关系和内容的融合分析,本文扩展 LDA 主题模型,构建了一个用于生成用户、标签、资源之间的关系以及由词汇组成的资源的内容的生成模型 TSM/Forc(tagging system model based on the fusion of object relation and content),以主题作为媒介将用户、标签、资源以及资源的内容关联起来,以便以概率的方式揭示四者之间的关系.

1.1 基础与假设

目前的标签推荐方法大多依靠对标签系统中对象之间的关系进行分析.通过将标签系统中用户、标签以及资源这 3 种对象之间的关系进行基于图的建模.这些方法发现形成这些关系的潜在原因,并利用这些信息实现推荐任务.这类推荐方法不考虑资源本身的特征,仅仅分析对象间的关系,在简化模型与分析过程的同时遗失了重要的资源内容特征.

本文在分析标签系统中对象间关系的同时,考虑资源内容的特征.进一步地,本文期望利用一个完整的模型融合地考虑两种信息,而不是分别考虑两种分析方法后将两种结果合并.为了更方便地阐述本文提出的模型,这里使用一组符号描述标签系统.

一个基本的标签系统中通常包含 3 种类型的对象:使用标签标记资源的用户、标签本身以及被标签标记的资源.这 3 种对象及其关系可以描述为一个三部图[3]:资源 $\boldsymbol{D}=\{d_1,\ldots,d_M\}$、用户 $\boldsymbol{U}=\{u_1,\ldots,u_U\}$、标签 $\boldsymbol{T}=\{t_1,\ldots,t_T\}$ 以及这些节点之间用以表示标注关系的超边集 $\boldsymbol{E}$.

为了使这一模型能够描述资源本身的特点,可以进一步地考虑每一个资源由一些词组成 $d_n=\boldsymbol{w}_n=\{w_{\langle 1\rangle},\ldots\}$.令 $\boldsymbol{W}=\{w_1,\ldots,w_W\}$表示由这些词所组成的不包含重复词汇的词表,则一个标签系统是一个五元组 $\boldsymbol{F}:=(\boldsymbol{U},\boldsymbol{T},\boldsymbol{D},\boldsymbol{E},\boldsymbol{W})$,如图 1 所示.图中带有相同编号的边组成了描述标注关系的超边,用户与资源之间的虚线则表示在仅能由资源创建者才能向资源附加标签的情况下隐含的用户与资源之间的创建关系.图中代表用户、标签与资源的符号互相区分,代表词汇的符号则可能代表相同的词汇.

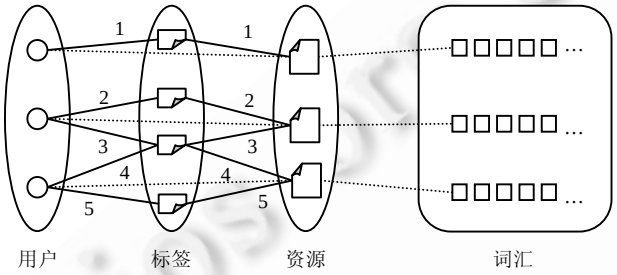

Fig.1 Tagging system model

图 1 标签系统模型

本文考虑使用基于概率方法的 LDA 主题模型作为融合表示关系与内容的基础.作为一个生成主题模型,LDA 在机器学习和自然语言理解等领域获得了广泛的关注.其完全生成模型的本质使得它可以很容易地扩展到新的文档和词汇上,并避免类似 Probabilistic Latent Semantic Indexing(pLSI)[7]模型中存在的过度拟合问题.

在 LDA 中,文档 d_n 中的每一个词 w_n 都被认为是根据一个潜在的主题 z_n 的多项分布得到的,而 z_n 则来自文档 d_n 主题的多项随机分布.依据这一过程,给定一个主题 k,观察到词表中的一个词 w 的概率表示为 $\varphi_{w|k}=P(w_n=w|z_n=k)$,而给定一个文档,观察到一个主题的概率表示为 $\theta_{k|d}=P(z_n=k|d_n=d)$.因此,给定所有的词汇以及主题的概率 Φ和 Θ,观察到一篇文章的概率是 $P(w\mid\Phi,\Theta)=\prod_{n=1}^{N}\sum_{k=1}^{K}\varphi_{w_n|k}\theta_{k|d_n}$.其中:N 是以词数计算的文章的长度;K 是潜在主题的数量;θ_d 与 φ_k 则被认为服从对称 Dirichlet 先验分布,其超参数为 α和 β.

LDA 模型通常难以进行直接推理,而更多的是使用变分推理[6]以及吉布斯抽样方法[3].吉布斯抽样是一个马尔可夫链蒙特卡洛方法.这种方法构建一个通过多次迭代缓慢收敛于目标分布的马尔可夫链来估计复杂概率分布的参数.对于 LDA,每一步吉布斯抽样都服从于 $P(z_n=k\mid\boldsymbol{z}_{-n},\boldsymbol{w})\propto\frac{N_{-n,k}^{(w_n)}+\beta}{N_{-n,k}^{(\cdot)}+V\beta}\frac{N_{-n,k}^{(d_n)}+\alpha}{N_{-n}^{(d_n)}+K\alpha}$,其中,$\boldsymbol{z}_{-n}$ 表示除当前主题指派位置以外所有其他词汇的主题,V 是词汇表中词的数量,$N_{-n,k}^{(w_n)}$ 是除当前主题指派位置外主题 k 被赋予词 w_n 的次数,$N_{-n,k}^{(\cdot)}$ 是除当前主题指派位置外主题 k 被赋予任意词汇的次数,$N_{-n,k}^{(d_n)}$ 是除当前主题指派位置外文档 d_n 中主题 k 出现的次数,$N_{-n}^{(d_n)}$ 是除当前主题指派所对应的词外文档 d_n 的词数.当抽样算法对所有文章的每一个词运行足够多次直至结果收敛时,通过对吉布斯抽样得到的结果进行抽样,参数 Φ和 Θ可以估计为 $\varphi_{w|k}=\frac{N_k^{(w)}+\beta}{N_k^{(\cdot)}+V\beta},\theta_{k|d}=\frac{N_k^{(d)}+\alpha}{N^{(d)}+T\alpha}$,即平滑后的主题 k 被指派给词汇 w 的次数除以主题 k 被指派的总次数以及平滑后的文档 d 中被指派为主题 k 的词汇数除以文档 d 中词汇的总数.

1.2 单用户情况下的TSM/Forc模型的描述

给定用户集合 $\boldsymbol{U}$、标签集合 $\boldsymbol{T}$、词表 $\boldsymbol{W}$、系统中资源的总数量 M 以及一组参数,本文考虑一个基于概率的生成过程用于生成标签系统 $\boldsymbol{F}$ 中的超边集 $\boldsymbol{E}$ 以及资源集合 $\boldsymbol{D}$.为了简化对问题的描述,这里首先假设标签系统中的资源仅可以由用户创建,同时仅可以由创建用户为资源附加标签(简称为单用户情况).对于一个标签系统 $\boldsymbol{F}$,本文假设其服从如下的初始化过程:

选择 $\psi\sim Dir(\delta),\varphi\sim Dir(\beta),\rho\sim Dir(\gamma)$.$Dir(\cdot)$是狄利克雷分布,$\delta,\beta$和$\gamma$是狄利克雷分布的超参数,$\psi,\varphi,\rho$分别是 $U\times K,W\times K,T\times K$ 维矩阵.给出了在不同的主题下,按多项分布随机地抽取到不同用户、词汇、标签的条件概率.

对于 $\boldsymbol{F}$ 中的一个资源 d_m,本文假设其服从如下生成过程:

1) 确定资源 d_m 内容的长度以及标签的数量:选择 $N\sim Poisson(\xi),P\sim Poisson(\varsigma)$.$N$ 给出了当前文档的长度,P 给出了文档标签的数量.参数ξ和ς并不需要被计算;
2) 确定资源 d_m 将会涉及的主题:选择$\theta\sim Dir(\alpha)$.θ给出了给定一篇文章,不同主题按多项分布出现的条件概率,α是狄利克雷分布的超参数;
3) 确定创作资源 d_m 的用户:
   a) 选择一个主题 $z^{(u)}\sim Multinomial(\theta)$,并且
   b) 依据 $u\sim Multinomial(\psi^{(z^{(u)})})$ 选择一个用户;
4) 确定资源 d_m 的内容:对于 N 个词中的每一个词 w_n:
   a) 选择一个主题 $z_n^{(w)}\sim Multinomial(\theta)$,并且
   b) 依据 $w_n\sim Multinomial(\varphi^{(z_n^{(w)})})$ 选择一个词;
5) 确定资源 d_m 所具有的标签:对于 P 个标签中的每一个标签 t_p:
   a) 选择一个主题 $z_n^{(t)}\sim Multinomial(\theta)$,并且
   b) 依据 $t_p\sim Multinomial(\rho^{(z_p^{(t)})})$ 选择一个标签,最后
   c) 将边(u,t_p,d_m)加入到超边集 $\boldsymbol{E}$ 中.

通过这一过程,可以生成每一个资源 d_m 的内容、确定的超边集 $\boldsymbol{E}$,并最终生成标签系统 $\boldsymbol{F}$.从这一生成过程中可以看到,资源的创建者、资源的内容以及资源所具有的标签选定都依赖于资源本身的主题分布.因此,以主题作为媒介,用户、标签、资源以及组成资源的词汇被关联起来,并使用一个统一的模型进行融合地描述.

根据这一过程,单用户情况下,TSM/Forc 模型可以形式化地概括如下:

$$\left.\begin{array}{rcc} u\mid z^{(u)},\psi^{(z^{(u)})} & \sim & Multinomial(\psi^{(z^{(u)})}) \\ w_n\mid z_n^{(w)},\varphi^{(z_n^{(w)})} & \sim & Multinomial(\varphi^{(z_n^{(w)})}) \\ t_p\mid z_p^{(t)},\rho^{(z_p^{(t)})} & \sim & Multinomial(\rho^{(z_p^{(t)})}) \\ \psi & \sim & Dir(\delta) \\ \varphi & \sim & Dir(\beta) \\ \rho & \sim & Dir(\gamma) \\ (\boldsymbol{z}^{(u)},\boldsymbol{z}^{(w)},\boldsymbol{z}^{(t)})=\boldsymbol{z}\mid\theta & \sim & Multinomial(\theta) \\ \theta & \sim & Dir(\alpha) \end{array}\right\} \tag{1}$$

1.3 多用户情况下的TSM/Forc模型的描述

考虑任意用户可以为任意资源添加标签的情况(称为多用户情况),则一篇文档的生成过程为:

1) 选择 $N\sim Poisson(\cdot),U_{d_m}\sim Poisson(\cdot),\boldsymbol{P}=\{P_1,...,P_{U_{d_m}}\}$,其中,$P_i\sim Poisson(\cdot),|\boldsymbol{P}|=P.N$ 给出了文档的长度,U_{d_m} 给出了为文档提供标签的用户数量,P_i 给出了用户 i 为文档添加的标签数量;
2) 选择$\theta\sim Dir(\alpha)$;
3) 对于 N 个词中的每一个词 w_n:

a) 选择一个主题 $z_n^{(w)} \sim Multinomial(\theta)$,并且

b) 依据 $w_n \sim Multinomial(\varphi^{(z_n^{(w)})})$ 选择一个词;

4) 对于 U_{d_m} 个用户中的每一个用户 u_i:

a) 选择一个主题 $z_n^{(u)} \sim Multinomial(\theta)$,并且

b) 依据 $u \sim Multinomial(\psi^{(z^{(u)})})$ 选择一个用户.而

c) 对于用户 u_i 所创建的 P_i 个标签中的每一个标签 t_p:

i) 选择一个主题 $z_p^{(t)} \sim Multinomial(\theta)$,并且

ii) 依据 $t_p \sim Multinomial(\rho^{(z_p^{(t)})})$ 选择一个标签,最后

iii) 将边(u_i,t_p,d_m)加入到超边集 $\boldsymbol{E}$ 中.

从上述生成过程可以看到,资源的每一个标签的生成过程仅依赖于资源本身的主题分布.因此,多用户情况下标签系统的生成过程也可以概括为系统(1)的形式,则系统(1)为 TSM/Forc 模型在任意情况下的概率描述.为了简化描述过程,本文接下来的部分将仅考虑单用户情况下的 TSM/Forc 模型.

1.4 TSM/Forc模型的直观解释

第 1.2 节给出的生成模型可以进行不严格但更加直观的解释.资源与标签系统中其他对象之间的关系以及资源的内容均可以视为资源本身的特征.可以认为,资源的所有特征,包括创建资源的用户、资源具有的标签以及组成资源的每一个词汇,均反映了资源所描述主题.例如,如果用户 *A* 经常创建信息技术领域的资源,则如果资源 *B* 创建自用户 *A*,则该资源很可能包含信息技术领域的主题.类似地,如果标签或者词汇 *C* 经常出现在信息技术领域的资源中,那么如果资源 *B* 包含词汇 *C*,则 *B* 也很有可能覆盖信息技术领域的主题.相反地,如果已知资源 *B* 覆盖信息技术领域的主题,那么资源 *B* 便有可能创建自用户 *A*,或者包含词汇或标签 *C*.在这些例子中,主题将资源与标签系统中其他的对象关联起来,使其可以直接关注资源的主题便可以容易地建立起资源与其他对象之间的关系.

假设标签系统 $\boldsymbol{F}$ 中存在一组潜在的主题 $\boldsymbol{K}=\{k_1,\ldots,k_K\}$,则对于一个资源 d_n,关心其所包含的不同主题的概率.进一步地,可以认为创作资源 d_n 的用户,资源被该用户所附加的每一个标签以及资源内容所包含的每一个词都服从给定资源主题的多项分布.这里,自然语言理解领域中经常使用的建模文章中词汇出现概率所使用的多项分布被进一步地应用到文章所具有的标签以及创建资源的用户上.因此,如果可以确定一个资源包含不同主题的概率,那么便可以依据在不同主题出现的条件概率下,某一个特定的词汇出现的概率计算出该资源中出现该词汇的概率.类似地,也可以计算出该资源被某一个标签标记或被某一个用户创作的概率.

基于这种解释,针对一个资源 d_n,首先依据标签系统 $\boldsymbol{F}$ 中不同主题的分布情况(由狄利克雷分布超参数α给出)确定 d_n 中不同主题出现的概率(即参数θ).接下来,分别依据给定主题的条件下用户出现的概率(即参数ψ)、词汇出现的概率(即参数φ)以及标签出现的概率(即参数ρ)为资源 d_n 指定用于构建标注关系的用户和标签对象,以及用于生成资源内容的所有词汇,便可以逐个资源地依据概率过程完成标签系统 $\boldsymbol{F}$ 的生成.

在实际应用中,需要计算模型中参数θ,ψ,φ与ρ才可以进行概率关系的定量计算.对这些参数的估计将在第 2.2 节中给出.

2 基于 TSM/Forc 模型的信息标签推荐方法

基于 TSM/Forc 模型,利用贝叶斯公式推导被计算对象之间的概率关系,可以计算标签系统中任意对象之间的条件概率关系用于推荐.利用吉布斯抽样方法可以确定模型中的未知参数并最终计算所需概率.

2.1 利用TSM/Forc模型进行标签推荐

本文提出的 TSM/Forc 模型继承了 LDA 的完全生成主题模型的特性,这使得利用该模型可以很容易地计算标签系统中不同对象之间的概率关系.利用吉布斯抽样方法确定模型参数后,对象之间的概率关系便可以确定.

通过进一步地利用贝叶斯公式,可以计算对象之间的关系.

给出一篇文档 d,这篇文档被用户 u 创作的概率可以表示为

$$P(u\mid d)=\sum_z P(u\mid z)P(z\mid d) \tag{2}$$

其中,$P(u|z)$即为 ψ_z^u ,$P(z|d)$即为 θ_z^d .该概率可以作为向用户推荐文章的依据.

当模型应用于标签推荐问题时,对于整体标签推荐问题,以及一篇文档 d,关心标签 t 出现的概率,即

$$P(t\mid d)=\sum_z P(t\mid z)P(z\mid d) \tag{3}$$

与公式(2)类似,公式(3)中所有的条件概率均可以使用模型的参数表示.针对文档 d,通过计算所有标签出现的概率即可完成整体标签推荐.而对于个性化标签推荐问题,关心用户 u 为文章 d 附加标签 t 的概率,即

$$P(t\mid d,u)=\sum_z P(t\mid z,u)P(z\mid d) \tag{4}$$

其中,$P(z|d)$已知.对于 $P(t|z,u)$,设 $P(t\mid z,u)=\rho_{z,u}^{(t)}$,其参数估计将在第 2.2 节给出.对于每一个用户文章对$\langle d,u\rangle$,为有的标签按照式(4)计算相应概率,则可以进行个性化标签推荐.上述标签推荐过程可以归纳如图 2 所示.

1) 给定标签系统 $\boldsymbol{F}=(\boldsymbol{U},\boldsymbol{T},\boldsymbol{D},\boldsymbol{E},\boldsymbol{W})$,潜在主题数量 K 以及超参数α,δ,β和γ;
2) 利用吉布斯抽样方法确定模型参数θ,ψ,φ,ρ以及用于特定目的的附加参数;
3) 确定被推荐对象的条件概率关系并计算条件概率;
4) 依据条件概率对对象进行排序;
5) 取 Top-N 结果进行推荐.

Fig.2 Tag recommendation method based on the combined analysis of relation and content

图 2 基于关系与内容融合分析的标签推荐方法

2.2 TSM/Forc模型参数的确定

本文采用吉布斯抽样方法确定模型参数θ,ψ,φ与ρ,这要求确定全条件概率分布函数的形式.由于

$$P(u,\boldsymbol{w},\boldsymbol{t},\boldsymbol{z})=P(u|z^{(u)})P(w|\boldsymbol{z}^{(w)})P(\boldsymbol{t}|\boldsymbol{z}^{(t)})P(\boldsymbol{z}) \tag{5}$$

根据公式(1)提供的条件概率分布,利用共轭分布的性质有:

$$P(u\mid z^{(u)})=\left(\frac{\Gamma(U\delta)}{\Gamma(\delta)^U}\right)^K\prod_{k=1}^K\frac{\prod_u\Gamma(n_k^{(u)}+\delta)}{\Gamma(n_k^{(\cdot,(u))}+U\delta)} \tag{6}$$

$$P(\boldsymbol{w}\mid\boldsymbol{z}^{(w)})=\left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^K\prod_{k=1}^K\frac{\prod_w\Gamma(n_k^{(w)}+\beta)}{\Gamma(n_k^{(\cdot,(w))}+W\beta)} \tag{7}$$

$$P(\boldsymbol{t}\mid\boldsymbol{z}^{(t)})=\left(\frac{\Gamma(T\gamma)}{\Gamma(\gamma)^T}\right)^K\prod_{k=1}^K\frac{\prod_t\Gamma(n_k^{(t)}+\gamma)}{\Gamma(n_k^{(\cdot,(t))}+T\gamma)} \tag{8}$$

$$P(\boldsymbol{z})=\left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^M\prod_{m=1}^M\frac{\prod_k\Gamma(n_k^{(m,u)}+n_k^{(m,w)}+n_k^{(m,t)}+\alpha)}{\Gamma(n_\cdot^{(m,u)}+n_\cdot^{(m,w)}+n_\cdot^{(m,t)}+K\alpha)} \tag{9}$$

其中,

- $n_k^{(u)},n_k^{(w)}$ 和 $n_k^{(t)}$ 分别代表主题 k 被指派给用户 u、词汇 w 和标签 t 的次数,$n_k^{(\cdot,(u))},n_k^{(\cdot,(w))}$ 和 $n_k^{(\cdot,(t))}$ 分别代表主题 k 被指派给任意用户、词汇和标签的次数;
- $n_k^{(m,u)},n_k^{(m,w)}$ 和 $n_k^{(m,t)}$ 分别代表主题 k 在文档 d_m 中被指派给用户 u、词汇 w 和标签 t 的次数;
- $n_\cdot^{(m,u)},n_\cdot^{(m,w)}$ 和 $n_\cdot^{(m,t)}$ 分别代表文档 d_m 中用户、词汇和标签的总数.

将公式(6)~公式(9)带入公式(5),可以得到吉布斯抽样所需的全概率公式.通过消去相同的项,可以得到每一个吉布斯抽样都服从于:

$$P(z_i=k\mid \boldsymbol{z}_{-i},u,\boldsymbol{w},\boldsymbol{t})\propto\begin{cases}\dfrac{1}{U}\dfrac{n_k^{(d_i,w)}+n_k^{(d_i,t)}+\alpha}{n_{\cdot}^{(d_i,w)}+n_{\cdot}^{(d_i,t)}+K\alpha}, & i=1\\ \dfrac{n_{-i,k}^{(w_i)}+\beta}{n_{-i,k}^{(w)}+W\beta}\dfrac{n_k^{(d_i,u)}+n_{-i,k}^{(d_i,w)}+n_k^{(d_i,t)}+\alpha}{n_{\cdot}^{(d_i,u)}+n_{-i,\cdot}^{(d_i,w)}+n_{\cdot}^{(d_i,t)}+K\alpha}, & 1<i\leqslant N+1\\ \dfrac{n_{-i,k}^{(t_i)}+\gamma}{n_{-i,k}^{(t)}+T\gamma}\dfrac{n_k^{(d_i,u)}+n_k^{(d_i,w)}+n_{-i,k}^{(d_i,t)}+\alpha}{n_{\cdot}^{(d_i,u)}+n_{\cdot}^{(d_i,w)}+n_{-i,\cdot}^{(d_i,t)}+K\alpha}, & N+1<i\leqslant N+P+1\end{cases}\tag{10}$$

其中,

- $n_k^{(d_i,u)}, n_k^{(d_i,w)}$ 和 $n_k^{(d_i,t)}$ 表示文档 d_i 中主题 k 被指派给任意用户、词汇和标签的次数;
- $n_{\cdot}^{(d_i,u)}, n_{\cdot}^{(d_i,w)}$ 和 $n_{\cdot}^{(d_i,t)}$ 表示文档 d_i 中用户、词汇和标签的总数;
- $n_{-i,k}^{(w_i)}$ 与 $n_{-i,k}^{(t_i)}$ 表示不计算当前指派位置的情况下,所有文档中主题 k 被指派给词汇 w_i 与标签 t_i 的次数;
- $n_{-i,\cdot}^{(d_i,w)}$ 与 $n_{-i,\cdot}^{(d_i,t)}$ 表示不计算当前指派位置的情况下,文档 d_i 中用户、词汇和标签的总数.

当吉布斯抽样过程运行足够长的时间后,利用收敛的吉布斯抽样结果,可以将模型的参数估计如下:

$$\hat{\psi}_j^{(u)}=\frac{n_j^{(u)}+\delta}{n_j^{(\cdot(u))}+U\delta},\hat{\varphi}_j^{(w)}=\frac{n_j^{(w)}+\beta}{n_j^{(\cdot(w))}+W\beta},\hat{\rho}_j^{(t)}=\frac{n_j^{(t)}+\gamma}{n_j^{(\cdot(t))}+T\gamma},\hat{\theta}_j^{(d)}=\frac{n_j^{(d,u)}+n_j^{(d,w)}+n_j^{(d,t)}+\alpha}{n_{\cdot}^{(d,u)}+n_{\cdot}^{(d,w)}+n_{\cdot}^{(d,t)}+W\alpha},$$

即,

- $\hat{\psi}_j^{(u)}$ 为主题 j 被指派给用户 u 的次数除以主题 j 被指派给所有用户的总次数;
- $\hat{\varphi}_j^{(w)}$ 为主题 j 被指派给词汇 w 的次数除以主题 j 被指派给所有词汇的总次数;
- $\hat{\rho}_j^{(t)}$ 为主题 j 被指派给标签 t 的次数除以主题 j 被指派给所有标签的总次数;
- $\hat{\theta}_j^{(d)}$ 为主题 j 被指派给文档 d 中的用户、标签以及词汇的次数除以文档 d 中用户、标签以及词汇的总数.

对于公式(4)中的 $\rho_{z_i,u}^{(t)}$,则可以通过修改参数ρ的估计为 $\hat{\rho}_{j,u}^{(t)}=\dfrac{n_{j,u}^{(t)}+\gamma}{n_{j,u}^{(\cdot(t))}+T\gamma}$ 近似地给出.其中,

- $n_{j,u}^{(t)}$ 表示只计算用户 u 所使用的标签的情况下,主题 j 被指派给标签 t 的次数;
- $n_{j,u}^{(\cdot(t))}$ 表示只计算用户 u 所使用的标签的情况下,主题 j 被指派给任意标签的次数.

2.3 TSM/Forc模型的增量更新

随着标签系统中信息的不断演进,新的用户、标签、资源以及词汇被引入到标签系统中.这要求对 TSM/Forc 模型进行更新,以便推荐结果可以反映这些变化.随着数据规模的不断膨胀,重复的重新运行吉布斯抽样过程的时间开销无疑将是巨大的,因此有必要采用增量式的更新方法.在线增量更新采用吉布斯抽样的 LDA 模型的主要方法包括 Online LDA 方法[8]、增量吉布斯抽样器[9]以及粒子滤波方法[9].由于粒子滤波方法能够以比较高效的方法实现,这里研究采用基于粒子滤波的方法在线增量更新 TSM/Forc 模型.这一方法依赖于对于每一个新对象 i(可能是用户、标签或词汇),其依赖于已经观察到的对象的主题抽样概率,即:

$$P(z_i=k\mid \boldsymbol{z}_{-i},u,\boldsymbol{w},\boldsymbol{t})\propto\begin{cases}\dfrac{n_{-i,k}^{(u_i)}+\delta}{n_{-i,k}^{(u)}+U\delta}\dfrac{n_{-i,k}^{(d_i,u)}+n_k^{(d_i,w)}+n_k^{(d_i,t)}+\alpha}{n_{-i,\cdot}^{(d_i,u)}+n_{\cdot}^{(d_i,w)}+n_{\cdot}^{(d_i,t)}+K\alpha}, & \text{if object i is a new user}\\ \dfrac{n_{-i,k}^{(w_i)}+\beta}{n_{-i,k}^{(w)}+W\beta}\dfrac{n_k^{(d_i,u)}+n_{-i,k}^{(d_i,w)}+n_k^{(d_i,t)}+\alpha}{n_{\cdot}^{(d_i,u)}+n_{-i,\cdot}^{(d_i,w)}+n_{\cdot}^{(d_i,t)}+K\alpha}, & \text{if object i is a new word,}\\ \dfrac{n_{-i,k}^{(t_i)}+\gamma}{n_{-i,k}^{(t)}+T\gamma}\dfrac{n_k^{(d_i,u)}+n_k^{(d_i,w)}+n_{-i,k}^{(d_i,t)}+\alpha}{n_{\cdot}^{(d_i,u)}+n_{\cdot}^{(d_i,w)}+n_{-i,\cdot}^{(d_i,t)}+K\alpha}, & \text{if object i is a new tag}\end{cases}$$

其中,$-i$ 代表不计算对象 i 的主题指派.粒子滤波方法还需要计算再生序列中主题的抽样概率,其分布服从:

$$P(z_j=k\mid \boldsymbol{z}_{-j},u,\boldsymbol{w},\boldsymbol{t})\propto\begin{cases}\dfrac{n_{-j,k}^{(u_j)}+\delta}{n_{-j,k}^{(u)}+U\delta}\dfrac{n_{-j,k}^{(d_j,u)}+n_k^{(d_j,w)}+n_k^{(d_j,t)}+\alpha}{n_{-j,\cdot}^{(d_j,u)}+n_\cdot^{(d_j,w)}+n_\cdot^{(d_j,t)}+K\alpha}, & \text{if object j is a new user}\\ \dfrac{n_{-j,k}^{(w_j)}+\beta}{n_{-j,k}^{(w)}+W\beta}\dfrac{n_k^{(d_j,u)}+n_{-j,k}^{(d_j,w)}+n_k^{(d_j,t)}+\alpha}{n_\cdot^{(d_j,u)}+n_{-j,\cdot}^{(d_j,w)}+n_\cdot^{(d_j,t)}+K\alpha}, & \text{if object j is a new word,}\\ \dfrac{n_{-j,k}^{(t_j)}+\gamma}{n_{-j,k}^{(t)}+T\gamma}\dfrac{n_k^{(d_j,u)}+n_k^{(d_j,w)}+n_{-j,k}^{(d_j,t)}+\alpha}{n_\cdot^{(d_j,u)}+n_\cdot^{(d_j,w)}+n_{-j,\cdot}^{(d_j,t)}+K\alpha}, & \text{if object j is a new tag}\end{cases}$$

其中,$-j$ 的意义与公式(10)中$-i$ 的意义相同,代表不计算当前位置的主题指派情况.

3 实验评估

3.1 实验设定

为了验证本文提出的标签推荐方法的效果,实验将本文提出的方法(以下称为“TSM/Forc 方法”)与标签推荐领域中主流以及最新的方法进行了对比.实验选用了 Hotho 等人提出的基于 PageRank 的 FolkRank 的方法[3]以及 Symeonidis 等人提出的基于张量的高维奇异值分解(higher-order singular value decomposition,简称 HOSVD)方法[4]的两个不同版本作为对比方法.所有的方法都使用 Java,Ruby 和 R[10]进行了实现.实现 TSM/Forc 方法的过程中大量采用了 Phan 等人[11]提供的使用 Java 语言编写的利用吉布斯抽样进行 LDA 参数估计的工具 JGibbLDA 的源代码.所有的实验运行于一台具有 3.17GHz Core 2 Duo 处理器、4GB 内存、运行 Ubuntu GNU/Linux 10.04 版本的 PC.不同算法的参数则依照原文中的说明进行了调校.

为了评估这些方法,实验使用了来自新浪博客的真实数据集.通过监视新浪博客的最近更新列表,下载程序间歇性地抓取了 2009 年 1 月~12 月期间创建的约 25 万条博客文章,并存储在数据库中.这些数据库记录包括了文章的标题、标签、作者、发布时间以及正文等信息.在这一数据集中,每篇文章平均拥有 5.27 个标签.由于在创建文章的过程中用户被要求必须将文章进行一次互斥的分类,因此这些文章已经被分类到 22 个互斥的分类中.

由于推荐结果的评估依赖人工进行,而这一过程的开销非常巨大,这里进一步地选择了一个比较小的数据集作为实验数据集.实验将数据集中的文章限制在 10 个主要的分类中,过滤掉了那些在数据集中拥有少于 50 篇博客文章的用户,并随机地抽取了大约 40 名用户与他们的约 1 600 篇文章.这些文章的正文和标签还进行了分词,去除一组常用词并递归地删除了使用次数过少的标签与词汇、拥有过少标签与词汇的文章以及拥有过少文章的用户,并最终得到一个包含 35 个用户、592 个独立标签以及 1 444 篇文章的数据集.

利用这一数据集,对象间的关系被提取并用于对比方法.对于 TSM/Forc 方法,为了降低计算量并加快吉布斯抽样过程的收敛速度,实验只使用文章正文中出现的前 400 个词作为资源内容的替代描述.参照 LDA 对超参数的选取,TSM/Forc 方法的超参数的选择为:α=50/K,δ=β=γ=0.1.

3.2 潜在主题的数量

TSM/Forc 方法的另一个重要参数为潜在主题的数量 K,其选择应该使基于这些参数的 TSM/Forc 模型生成数据集的可能性取得最大.依据文献[12],K 应最大化 $P(u,\boldsymbol{w},\boldsymbol{t}|\boldsymbol{z},K)$的调和平均数,其结果如图 3 所示.

对于每一个 K 值,实验运行 8 条马尔可夫链,并忽略了前 1 000 次抽样结果.利用 50 次随机选取的抽样结果,计算 $P(u,\boldsymbol{w},\boldsymbol{t}|K)$并记录.从图中可以看到,当 K 取 30 时,$P(u,\boldsymbol{w},\boldsymbol{t}|K)$取得最大值.可以大致上认为,这一结果表明了实验数据集中大约包含 30 种不同的潜在主题.相比构建数据集时仅仅选定了属于 10 个主要分类的文章,TSM/Forc 方法可以提供比文章自身分类更加细致的描述能力.

表 1 给出了使用 TSM/Forc 方法分析数据集后,针对潜在主题 A 和潜在主题 B,具有最高出现概率的 10 个词汇和标签.从这些词汇与标签可以看到,主题 A 与 B 均与汽车相关,并且这一特征反映在其标签统计结果的第一项均为“汽车”.相比之下,主题 A 更侧重于技术特点,表现在词汇的使用上为“引擎”、“研发”等词汇与标签的

出现;而主题 B 则更侧重于市场信息,如“价格”、“市场”等词语的使用.

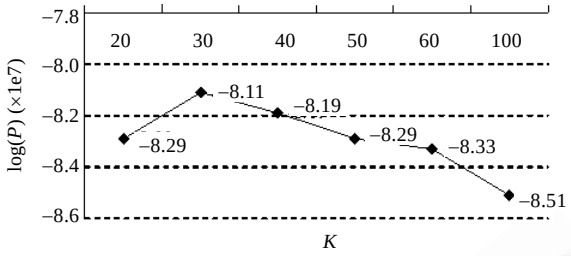

Fig.3 Estimated values of $P(u,\boldsymbol{w},\boldsymbol{t}|K)$ with different K

图 3 不同 K 值下,$P(u,\boldsymbol{w},\boldsymbol{t}|K)$的估计值

Table 1 Some of the words and tags with high probabilistic values of the TSM/Forc method

表 1 TSM/Forc 方法的部分高概率词汇和标签

	潜在主题 A	潜在主题 B
词汇	车型 动力 发动机 改装 引擎 系统 设计 全新 采用 包括	汽车 市场 车型 价格 产品 品牌 销量 企业 上市 销售
标签	汽车 改装 行者 中心 奔驰 研发 发现 卫士 跑车 宝马	汽车 焦点 评论 晓程 丹东 车型 高尔夫 购车 价格 上海

3.3 判断用户生成文章的能力

接下来考察数据集中用户生成每一篇文章的能力.第 2.1 节中,用户生成文章的能力可以通过计算 $P(u|d)=\sum_z P(u|z)P(z|d)$ 得到.这里对数据集中的每一个用户-文章对(u,d)计算 $P(u|d)$,并将关于某一用户 A 的计算结果如图 4 所示.

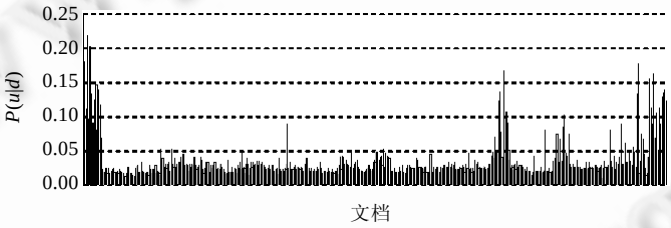

Fig.4 Probabilistic values of articles are created by a user A

图 4 所有文章被某用户 A 所创作的概率

图 4 中,黑色数据表示由用户 A 所创建的文章,灰色数据表示由其他用户创建的文章.可以明显地观察到,TSM/Forc 方法的计算结果认为,用户 A 基本上有相对较大的机会创建出那些确实由自己创建的文章.绝大多数其他用户创建的文章被认为难以被用户 A 创建,但用户 A 仍旧被认为很有可能有能力创建一些原本不是由他创建的文章.从图 4 中可以直观地看到,“用户 A 是否有能力创建一篇文章”这一标准的大致阈值在 0.05 附近.

为了量化地描述用户创建文章的能力阈值,这里认为 $P(u|d)$服从 Beta 分布.Beta 分布经常被用于作为随机变量的先验分布.包括 LDA 的 Dirichlet 超参数在内的很多随机变量都被假设服从 Beta 分布.这里使用那些确实由用户创建的文章作为训练样本,则 Beta 分布的参数被估计为: $\alpha=\bar{x}\left(\frac{\bar{x}(1-\bar{x})}{v}-1\right)$, $\beta=(1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{v}-1\right)$,其中, $\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i$ 为样本的均值, $v=\frac{1}{n}\sum_{i=1}^{N}(x_i-\bar{x})^2$ 为样本方差.使用 90%置信区间,可以得到用户 A 创建文章能力的阈值约为 0.054.可以看到,这一估计值与直观的感受一致.

3.4 标签推荐效果对比

为了简化对比不同方法的标签推荐结果,这里要求各个方法给出对于一个资源的最终标签推荐结果,而不是对于该资源,针对每一个不同的用户给出不同的个性化推荐结果.针对一个资源,所有的推荐结果被不区分地合并到在一起,并交予人工进行评价.其准确率与召回率定义为 $precision=TP/(TP+FP)$,$recall=TP/(TP+FN)$,其中,TP 为算法与人工评价均认为应该具有的标签数量,FP 为算法认为应该具有但人工评价认为不应该具有的标签数量,FN 为算法认为不应该具有但人工认为应该具有的标签数量.准确率与召回率还将被进一步地合并为 $F1$ 指标 $F1=(2\times precision\times recall)/(precision+recall)$,以便更加直观地综合反映方法的性能.

作为标签推荐的对比方法,实验选用了 FolkRank 方法以及 HOSVD 方法的两个不同版本.对于 FolkRank 方法,其阻尼系数 d 被设置为 0.7,其结果被迭代地计算,直到迭代结果之间的差值小于 10^{-6}.对偏好向量的计算依照原文进行,对应项权重被设置为 $1+|U|$与 $1+|R|$.计算的结果用 FolkRank 代表.对于 HOSVD 方法,实验对比了直接应用 HOSVD 方法(记为 HOSVD-Direct)以及使用高斯核函数进行平滑后的 HOSVD 方法(记为 HOSVD-Smoothed).所有的推荐结果被进一步地合并,以便得到最终推荐结果.

由于上述 3 种方法都基于个性化的推荐过程,这里也采用基于个性化标签推荐的 TSM/Forc 方法作为对比.由于

$$P(t\mid d)=\sum_u P(t,u\mid d)=\sum_u P(t\mid u,d)P(u\mid d) \tag{11}$$

则利用上式可以计算在个性化标签推荐结果的基础上给出最终推荐结果.该方法的推荐结果记为 Fusion-Direct.实验进一步地只考虑合并那些被认为有能力生成一篇文档的用户的推荐结果.即对于公式(11),要求

$$P(t\mid u,d)P(u\mid d)=\begin{cases}0, & \text{if } P(u\mid d)<\mu\\ P(t\mid u,d)P(u\mid d), & \text{if } P(u\mid d)\geqslant\mu\end{cases},$$

其中,μ为用户生成一篇文档的能力阈值,其值的计算见第 3.3 节.这种方法被记为 Fusion-Cut.

所有方法推荐结果的准确率、召回率和 $F1$ 指标呈现如图 5 与表 2 所示.从实验结果可以看出,相比当前主流和最新的标签推荐方法,TSM/Forc 方法可以提供更好的标签推荐结果.同时,考虑用户生成文章的能力,可以进一步地提升 TSM/Forc 方法的效果.

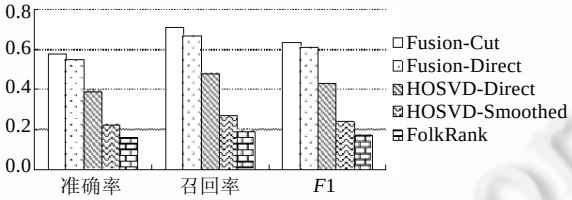

Fig.5 Comparison of tag recommendation results

图 5 标签推荐效果对比

Table 2 Tag recommendation results

表 2 标签推荐结果

	准确率	召回率	F1
Fusion-Cut	**0.58**	**0.71**	**0.64**
Fusion-Direct	0.55	0.67	0.61
HOSVD-Direct	0.39	0.48	0.43
HOSVD-Smoothed	0.22	0.27	0.24
FolkRank	0.16	0.19	0.17

3.5 TSM/Forc方法的复杂度分析

计算 TSM/Forc 方法时,其过程与计算采用吉布斯抽样方法的 LDA 类似,具有比较好的空间复杂度,其主要空间开销来自于抽样结果存储,并且可以很容易地计算出来.使用第 1.2 节的符号,假设每篇文档的长度最大为

N_{max},标签数最多为P_{max},则空间开销≤$M\times(N_{max}+P_{max}+1)$.为了提升吉布斯抽样速度,可以进一步地设置缓存变量记录抽样过程.其主要空间开销来自于缓存矩阵,具体为$K\times(W+U+T)$.其余缓存向量以及计算时空间开销相比前二者则可以忽略.

与空间复杂度相同,TSM/Forc 方法同样具有类似 LDA 的时间复杂度.在每一轮抽样过程中,由于抽样过程需要更新每一个抽样结果,其时间复杂度同样≤$M\times(N_{max}+P_{max}+1)$.影响时间复杂度的另一个因素是抽样需要迭代进行的次数.在实验中,实验结果通常在几百次迭代后即收敛.文献[13]介绍了判断马尔可夫链蒙特卡洛方法结果收敛的方法.

4 相关工作

在标签系统中,用户以关键字的形式向共享资源添加元数据.很多研究工作都揭示了标签系统的各种特点.Golder 等人[14]分析了标签系统的结构及其动态特征,并发现标签具有极其丰富的用法.依照用法的不同,标签可以分为识别资源主题的名词性标签,如“新闻”;识别资源类别的标签,如“图书”;识别资源特征的标签,如“有趣”;识别个性化分类的标签,如“我的文章”以及自我描述性的标签,如“待阅读”等等.这些多样化的标签使用方法使标签可以用于任何信息的分类,同时也增加了利用标签信息进行自动化信息处理的难度.Halpin 等人[15]给出了一个标签系统的生成模型,以便进一步研究其动态特性.他们认为,对于一个资源来说,其热门标签不会随时间改变,并且标签的使用服从幂率分布.通常,研究工作关注于标签系统中的 3 种对象:用户、标签以及资源.

标签推荐领域的相关研究在很大程度上受到搜索引擎领域研究的影响.经典的链接分析方法如 HITS[16]与 PageRank[17]在标签推荐领域中都有应用.Xu 等人[18]提出的 PR 算法利用类似 HITS 算法的过程为每个用户指定一个权威指标,以便描述用户过去的标记行为.这种方法使得推荐结果可以获得所期望的一些较好的性质,如结果通常可以覆盖多个层面的信息,并获得较高的召回率,同时还能使推荐结果具备较高的使用频率以确保其质量.Hotho 等人[3]提出的 FolkRank 算法则模拟用户在资源正文与用户信息等页面之间的跳转过程,并采用类似 PageRank 的计算方法.其核心思想在于,被重要的用户使用重要的标签标注的资源,其自身也是重要的.这一关系也可以对称地应用于用户以及标签.利用这种相互增强关系可以构建类似 PageRank 的马尔可夫过程,并计算标签推荐的结果.近年来,随着高维异构对象数据挖掘技术的发展,标签推荐领域中也出现了采用相关技术的方法.如 Xu 等人[19]与 Symeonidis 等人[4]均提出了基于高维奇异值分解(HOSVD)[20]的标签推荐方法.通过将矩阵 SVD 扩展到张量空间中,HOSVD 可以获得近似于矩阵 SVD 的特性,并揭示多个异构对象之间的潜在关系.利用这种方法,可以比较容易地建模标签系统中的多种不同对象,捕捉比上述方法更加全面的信息.文献[4]将 HOSVD 方法与最近的方法进行了比较,并提出了一个可以用于标签系统中多种不同推荐任务的统一框架.

与这些方法不同,本文提出了一种基于 LDA 的、可以融合描述标签系统中对象间关系以及资源内容的模型,使推荐方法不再局限于分析对象间关系而可以扩展到资源本身的特征.一些研究方法也将 LDA 应用于标签推荐领域.Harvey 等人[5]提出的 TTM 模型通过扩展 LDA 模型,完整地描述了标签系统中用户、标签与资源的三部结构,使其可以估计用户与文档的主题分布以及主题的词汇分布.Xu 等人[21]则在考虑标签的 LDA 生成过程的同时,考虑了资源之间的相似性,将资源之间相似性的产生原因也建模为依条件概率的随机过程,并在此基础之上提出了 Regularized LDA 模型.这些模型的提出,也表明了 LDA 具有极好的可扩展性,并可以应用于多种类型的任务.

一些研究工作也致力于不同领域中的对象关系与对象内容的融合分析.如 Shakery 等人[22]提出的相关性传播框架,利用概率方法将内容与链接信息合并起来,以提供更好的搜索结果.这种方法可以以一种标准的方法全面地利用内容信息以及链接结构信息,并可以使用多数已有的链接算法.通过计算基于内容的相关性概率,这一方法将相关性在不同类型的邻居节点间传播,实现了对不同信息的融合利用.Calado 等人[23]利用贝叶斯网络将利用链接分析与文本分析的结果合并起来,用于分类 Web 文本.这种方法提供了独立于链接算法与内容算法的合并方法,使其可以很容易地合并任意类型的计算结果.通过融合方法,可以提供优于单一方法的计算结果,对融合模型的研究也正逐渐成为相关领域的研究热点.

5 结论与未来研究

本文提出了一种新的概率潜在主题模型 TSM/Forc,以便融合地建模并分析标签系统中用户、标签与资源之间的关系以及资源本身的内容.这一模型可以融合统一地描述对象间的关系以及资源本身的特征,并可以很容易地以条件概率的方式描述、建模和计算标签系统中任意对象之间的关联.基于这一模型,本文提出了用于 Web 2.0 资源标签推荐的融合方法,并使用来自真实应用的数据集验证了这一方法.实验结果表明,相比当前主流与最新的标签推荐方法,本文提出的融合方法可以获得更好的标签推荐效果.

由于本文提出了标签系统的完全模型,因此在未来的研究中,可以很容易地将这一模型应用于标签推荐以外的多种任务中.例如,这一模型可以用于寻找具有类似兴趣的用户以便进行好友推荐,或者用于对类似的资源进行资源推荐.由于模型考虑了组成资源的词汇,因此其还可以很容易地用于改善搜索引擎结果的排序.进一步地,模型提供了对标签系统中各种对象与资源的抽象描述能力,因此还可以应用于个性化搜索与用户建模等领域.

References:

[1] Guy M, Tonkin E. Folksonomies: Tidying up tags? D-Lib Magazine, 2006,12(1). [doi: 10.1045/january2006-guy]

[2] Sigurbjörnsson B, van Zwol R. Flickr tag recommendation based on collective knowledge. In: Huai JP, Chen R, Hon HW, Liu YH, Ma WY, Tomkins A, Zhang XD, eds. Proc. of the 17th Int'l Conf. on World Wide Web. New York: ACM, 2008. 327−336. [doi: 10.1145/1367497.1367542]

[3] Hotho A, Jäschke R, Schmitz C, Stumme G. Information retrieval in folksonomies: Search and ranking. In: Sure Y, Domingue J, eds. Proc. of the Semantic Web: Research and Applications, 3rd European Semantic Web Conf. Heidelberg: Springer-Verlag, 2006. 411−426. [doi: 10.1007/11762256_31]

[4] Symeonidis P, Nanopoulos A, Manolopoulos Y. A unified framework for providing recommendations in social tagging systems based on ternary semantic analysis. IEEE Trans. on Knowledge and Data Engineering, 2010,22(2):179−192. [doi: 10.1109/TKDE.2009.85]

[5] Harvey M, Baillie M, Ruthven I, Carman M. Tripartite hidden topic models for personalised tag suggestion. In: Gurrin C, He YL, Kazai G, Kruschwitz U, Little S, Roelleke T, Rüger SM, van Rijsbergen K, eds. Advances in Information Retrieval, the 32nd European Conf. on IR Research. Heidelberg: Springer-Verlag, 2010. 432−443. [doi: 10.1007/978-3-642-12275-0_38]

[6] Blei DM, Ng AY, Jordan MJ. Latent dirichlet allocation. Journal of Machine Learning Research, 2003,3:993−1022. [doi: 10.1162/jmlr.2003.3.4-5.993]

[7] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 2001,42(1/2):177−196. [doi: 10.1023/A:1007617005950]

[8] Banerjee A, Basu S. Topic models over text streams: a study of batch and online unsupervised learning. In: Proc. of the 2007 SIAM Int'l Conf. on Data Mining. Philadelphia: SIAM, 2007. 431−436.

[9] Canini KR, Shi L, Griffiths TL. Online inference of topics with latent dirichlet allocation. In: van Dyk D, Welling M, eds. Proc. of the 12th Int'l Conf. on Artificial Intelligence and Statistics. Boston: MIT Press, 2009. 65−72.

[10] R Development Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2010. http://www.r-project.org/

[11] Phan XH, Nguyen LM, Horiguchi S. Learning to classify short and sparse text & Web with hidden topics from large-scale data collections. In: Huai JP, Chen R, Hon HW, Liu Y, Ma WY, Tomkins A, Zhang XD, eds. Proc. of the 17th Int'l Conf. on World Wide Web. New York: ACM, 2008. 91−100. [doi: 10.1145/1367497.1367510]

[12] Griffiths TL, Steyvers M. Finding scientific topics. Proc. of the National Academy of Sciences of the United States of America, 2004,101(Suppl.):5228−5235. [doi: 10.1073/pnas.0307752101]

[13] Liu JS. Monte Carlo Strategies in Scientific Computing. New York: Springer Science+Business Media, LLC, 2001.

[14] Golder SA, Huberman BA. The Structure of Collaborative Tagging Systems. arXiv:cs/0508082v1:Ithaca: Cornell University Library, 2005. http://arxiv.org/pdf/cs/0508082v1

[15] Halpin H, Robu V, Shepherd H. The complex dynamics of collaborative tagging. In: Williamson CL, Zurko ME, Patel-Schneider PF, Shenoy PJ, eds. Proc. of the 16th Int'l Conf. on World Wide Web. New York: ACM, 2007. 211−220. [doi: 10.1145/1242572.1242602]

[16] Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of ACM, 1999,46(5):604−632. [doi: 10.1145/324133.324140]

[17] Page L, Brin S, Motwani R, Winograd T. The Pagerank Citation Ranking: Bringing Order to the Web. SIDL-WP-1999-0120: Stanford: Stanford Digital Library, 1999. http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

[18] Xu Z, Fu Y, Mao JC, Su DF. Towards the semantic Web: Collaborative tag suggestions. In: Proc. of the Collaborative Web Tagging Workshop at 15th Int'l World Wide Web Conf. 2006.

[19] Xu YF, Zhang L, Liu W. Cubic analysis of social bookmarking for personalized recommendation. In: Zhou XF, Li JZ, Shen HT, Kitsuregawa M, Zhang YC, eds. Frontiers of WWW Research and Development—APWeb 2006. Heidelberg: Springer-Verlag, 2006. 733−738. [doi: 10.1007/11610113_66]

[20] Lathauwer LD, Moor BD, Vandewalle J. A multilinear singular value decomposition. SIAM Journal of Matrix Analysis and Applications, 2000,21(4):1253−1278. [doi: 10.1137/S0895479896305696]

[21] Xu H, Wang JD, Hua XS, Li SP. Tag refinement by regularized LDA. In: Gao W, Rui Y, Hanjalic A, Xu CS, Steinbach EG, El-Saddik A, Zhou MX, eds. Proc. of the 17th Int'l Conf. on Multimedia 2009. New York: ACM, 2009. 573−576. [doi: 10.1145/1631272.1631359]

[22] Shakery A, Zhai CX. A probabilistic relevance propagation model for hypertext retrieval. In: Yu PS, Tsotras VJ, Fox EA, Liu B, eds. Proc. of the 2006 ACM CIKM Int'l Conf. on Information and Knowledge Management. New York: ACM, 2006. 550−558. [doi: 10.1145/1183614.1183693]

[23] Calado P, Cristo M, Moura E, Ziviani N, Ribeiro-Neto B, Gonçalves MA. Combining link-based and content-based methods for Web document classification. In: Proc. of the 2003 ACM CIKM Int'l Conf. on Information and Knowledge Management. New York: ACM, 2003. 394−401. [doi: 10.1145/956863.956938]

张斌(1964－),男,辽宁开原人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为 Web 信息处理,服务计算,数据挖掘.

张引(1985－),男,博士生,主要研究领域为 Web 信息处理.

高克宁(1963－),女,博士,教授,主要研究领域为 Web 信息处理.

郭朋伟(1984－),男,博士生,主要研究领域为 Web 信息处理.

孙达明(1981－),男,博士生,CCF 学生会员,主要研究领域为 Web 信息集成.