文章编号: 1007-5321(2014)03-0038-05 DOI: 10.13190/j.jbupt.2014.03.008

一种用于社会化标签推荐的主题模型

孙甲申， 王小捷
(北京邮电大学 计算机学院，北京 100876)

摘要: 社会化标签中普遍存在标签的主题粒度和文档不一致以及部分标签和文档内容无关这两个问题，而现有基于主题模型的社会化标签推荐算法并没有同时对二者进行建模. 针对这两点，提出了一种新的主题模型，该模型不仅允许标签和文档具有各自的主题粒度，而且允许标签来自与文档无关的噪声主题. 在两个不同的社会化标签语料上的实验结果表明，所提出的模型相比内容相关模型和标签的隐含狄利克雷分配模型，在混淆度和平均正确率均值这两个指标上均有所提高.

关 键 词: 社会化标签推荐; 主题模型; 标签主题粒度; 噪声标签
中图分类号: TN929.53 **文献标志码**: A

A Topic Model for Social Tag Recommendation

SUN Jia-shen， WANG Xiao-jie
(School of Computer Science，Beijing University of Posts and Telecommunications，Beijing 100876，China)

Abstract: It is common that the topic-granularity of social tags is not consistent with correspondent document，and some tags cannot describe the topic of the document content. The existing topic models-based tag recommendation did not address the foregoing problems simultaneously as well. Motivated by the fact，the proposed novel topic model allows different granularity of word topics and tag topics，and assumes that the tags can originate from a general distribution unrelated to the content. Experimental results show that the proposed model outperforms content relevance model (CRM) and tag-logical device address (tag-LDA) on two different social tagging corpora in both perplexity and mean average precision.

Key words: social tag recommendation; topic model; tag-granularity; noisy tags

社会化标签允许用户用自造的标签来标注网络资源，对标签的内容、个数和一致性均无限制，因此较容易被互联网用户接受. 但是，由于网络资源非常巨大，人工标注费时费力，难以实施，同时，人工标注中也存在大量错标、标注不一致的情况，这对标签的实用性带来了很大的困难. 因此，利用计算机进行自动、高质量的标签推荐，一直以来都是业内学者研究的热点.

与基于协同过滤的方法不同，本研究是基于文档内容的社会化标签推荐方法. 该类方法[1-2]可以不受新文档和冷门话题的限制，更适用于对有充足文本内容的文档进行推荐，如网页、博客文章、新闻等. 而随着隐含狄利克雷分配模型(LDA，latent Dirichlet allocation)[3]的发展，越来越多的研究工作将 LDA 引入标签推荐任务[2,4-7]. 虽然，这些主题模型一定程度上解决了对有标签语料的建模问题，但是，由于社会化标签的自由性，对比传统的标签，有其自有的特点: 1) 社会化标签和对应文档的主题粒度差异较大; 2) 社会化标签中含有大量与被标注内容无关的噪声标签. 这其中包括用来描述内容以外

收稿日期: 2013-12-01
基金项目: 国家自然科学基金项目(61273365); 国家高技术研究发展计划项目(2012AA011104)
作者简介: 孙甲申(1984—)，男，博士，E-mail: b.bigart911@gmail.com; 王小捷(1969—)，男，教授，博士生导师.

的标签,以及用户有意或无意使用了错误的标签.

针对这两点,本研究将标签和文档可生成自不同粒度主题的思想以及标签中存在噪声主题的现象进行统一建模,提出了一个新的主题模型: 引入标签粒度和噪声的 LDA(TN-LDA, tag-granularity and noise-aware LDA),并基于该主题模型对未见文档进行标签推荐.

1　基于标签粒度及噪声标签的模型

1.1　TN-LDA 模型

TN-LDA 模型的生成过程如下,对应的生成图如图 1 所示.

1) 抽样相关概率 λ,服从 beta(γ)

2) 对每个词主题 $k \in 1,2,\cdots,K$,抽样一个词层的分布 φ_w^k,服从 dir(β_w); 抽样一个标签主题(子主题)层的分布 Ψ_k,服从 dir(η)

3) 对每个标签主题 $q \in 0,1,\cdots,K_1$,抽样一个标签层的分布 φ_t^q,服从 dir(β_t),其中 $q=0$ 表示与文档内容无关的噪声主题,$q \in 1,2,\cdots,K_1$ 表示与文档内容相关的标签主题

4) 对应每个文档 $d \in 1,2,\cdots,D$,抽样一个主题层的分布 θ_d,服从 dir(α)

5) 对该文档 d 的每个词 $w_{dn} \in 1,2,\cdots,N$

①基于多项式分布 mult(θ_d),抽样一个词主题 $z_{dn} \in 1,2,\cdots,K$

②基于当前主题 z_{dn} 决定的多项式分布 mult($\varphi_w^{z_{dn}}$),抽样一个词 w_{dn}

6) 对该文档 d 的每个标签 $t_{dm} \in 1,2,\cdots,M$

①基于多项式分布 mult(θ_d),随机抽取一个词主题 $s_{dm} \in 1,2,\cdots,K$

②基于当前主题 s_{dm} 决定的多项式分布 mult($\Psi^{s_{dm}}$),抽样一个标签主题 $y_{dm} \in 1,2,\cdots,K_1$

③基于二项式分布 Bernoulli(λ),抽取 r_{dm}

a. 若 $r_{dm}=1$,基于当前标签主题 y_{dm} 决定的多项式分布 mult($\varphi_t^{y_{dm}}$),抽样一个标签 t_{dm}

b. $r_{dm}=0$,基于噪声主题的多项式分布 mult(φ_t^0),抽样一个标签 t_{dm}

1.2　基于吉布斯采样的模型参数估计

用吉布斯抽样方法对 TN-LDA 模型进行参数估计. 给定文档的词 $W=\{w_d\}_{d=1}^D$、标签 $T=\{t_d\}_{d=1}^D$、词的主题 $Z=\{z_d\}_{d=1}^D$、标签的主题 $S=\{s_d\}_{d=1}^D$ 和 $Y=\{y_d\}_{d=1}^D$,以及相关量 $R=\{r_d\}_{d=1}^D$,当前文档 d 中,

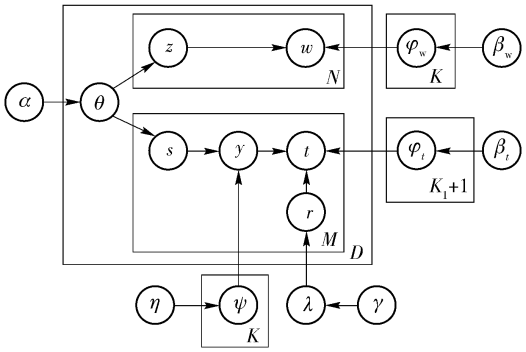

图 1　TN-LDA 模型的生成过程

第 i 个词语 w_i 属于主题 k 的概率为

$$p(z_i=k \mid Z_{-i},W,T,S,Y,R,\Theta) \propto \frac{N_{kd}^{-i}+\alpha}{N_d^{-i}+K\alpha} \cdot \frac{N_{kw_i}^{-i}+\beta_w}{N_k^{-i}+W\beta_w} \tag{1}$$

其中: 上标 $-i$ 为不考虑当前词 i 的计数; N_{kd}^{-i} 为当前文档 d(包括词语和标签)中含有主题 k 的频次; N_d^{-i} 为当前文档 d 中词语和标签总数; $N_{kw_i}^{-i}$ 为在整个文档集中,词 i 和主题 k 的共现次数; N_k^{-i} 为在整个文档集中主题 k 的出现次数; K 为词主题的个数; W 为文档集中词表的大小; α 和 β_w 分别为 Dirichlet 分布的超参数.

当前文档 d 中,第 j 个标签 t_j 和文档相关的情况下,属于主题 k 以及标签主题 q 的概率为

$$p(s_j=k,y_j=q,r_j=1 \mid S_{-j},Y_{-j},R_{-j},Z,W,T,\Theta) \propto \frac{N_{kd}^{-j}+\alpha}{N_d^{-j}+K\alpha} \cdot \frac{N_{kq}^{-j}+\eta}{N_k^{-j}+K_1\eta} \cdot \frac{M_{qt_j}^{-j}+\beta_t}{M_q^{-j}+T\beta_t} \cdot \frac{M^{-j}-M_0^{-j}+\gamma}{M^{-j}+2\gamma} \tag{2}$$

其中: 上标 $-j$ 为不考虑当前标签 j 的计数,与式(1)中仅有上标不同的变量将不作单独说明; N_{kq}^{-j} 表示主题 k 和标签主题 q 的所有共现次数; N_k^{-j} 表示在整个文档集中主题 k 的出现次数; $M_{qt_j}^{-j}$ 表示在整个文档集中,标签 j 和标签主题 q 的共现次数; M_q^{-j} 表示在所有标签中标签主题 q 的出现次数; M^{-j} 和 M_0^{-j} 分别表示所有标签数和噪声标签数; K_1 表示标签主题的个数; T 表示文档集中标签表的大小; β_t 和 η 分别为 Dirichlet 分布的超参数; γ 为 beta 分布的超参数.

第 j 个标签 t_j 和文档不相关,属于噪声主题的概率为

$$p(y_j=0,\ r_j=0 \mid S_{-j},Y_{-j},R_{-j},Z,W,T,\Theta) \propto \frac{M_{0t_j}^{-j}+\beta_t}{M_0^{-j}+T\beta_t} \cdot \frac{M_0^{-j}+\gamma}{M^{-j}+2\gamma} \tag{3}$$

其中: $M_{0t_j}^{-j}$表示在整个文档集中,标签 j 和噪声主题"0"的共现次数.

根据上述吉布斯抽样过程的结果,TN-LDA 模型可以得到更新后的模型参数,并对未见文档 d_{new} 进行推导. 然后根据式(4)计算训练集标签的权值,并将满足式(4)的前若干个标签推荐为该文档的标签.

$$t_d = \arg\max \sum_{q=1}^{K_1} p(t \mid y_q)\, p(y_q \mid d_{\text{new}}) \tag{4}$$

其中,更新后的标签主题-标签分布为

$$p(t_j \mid y_q) = \hat{\beta}_t = \frac{M_{qt_j} + \beta_t}{M_q + T\beta_t} \tag{5}$$

更新后的文档-标签主题分布为

$$p(y_q \mid d) = \hat{\theta}_1 = \frac{N_{qd} + \alpha}{N_d + K_1\alpha} \tag{6}$$

其中 N_{qd}表示当前文档 d(只包括标签)中含有标签主题 q 的频次.

2 实验及分析

2.1 数据准备

实验采用2个不同性质的社会化标签数据集,其中 BIBTEX 来自学术论文收藏网站 Bibsonomy①,BIBTEX 数据集中的文档长度很短,主题相对集中在计算机科学和生物科学等研究领域,标签多为描述较细粒度的专业术语等概念词,极少含有噪声标签.

第2个数据集 Delicious 来自于分布式人工智能(DAI,distributed artificial intelligence)实验室的 Wetzker 等[8]提供一份从 Delicious② 上抓取的数据: dai_labor_delicious.

2.2 混淆度比较

在主题模型中,混淆度(perplexity)被用来评测模型是否较好的拟合数据,混淆度越低表示更好的生成性能. 混淆度的定义为

$$P_t = \exp\left[\frac{-\sum_d \sum_{t \in d} \ln p(t)}{\sum_d (N_d^w + N_d^t)}\right] \tag{7}$$

图2和图3分别为两个数据集下,TN-LDA 模型在不同的词主题数和标签主题数组合下的混淆度变化以及和内容相关模型(CRM,content relevance model)、Tag-LDA 的比较结果. 其中,图示表示对应的模型,括号内的数值表示 TN-LDA 模型的词主题数.

首先,由图2可见,一方面,当词主题数为200时,TN-LDA 模型的混淆度最低,其中标签主题数为800时混淆度达到最低值,这说明该模型的词主题数和标签主题数分别为200和800时,最适合对 BIBTEX 建模. 另一方面,Tag-LDA 和 CRM 的混淆度随主题数的增加而增加,说明这两个模型过拟合 BIBTEX 数据. 而 TN-LDA 的混淆度趋势则相反,且在词主题数为200时基本低于 Tag-LDA 的混淆度,说明 TN-LDA 相比其他2个模型,当采用适当的主题数时,更适合对 BIBTEX 建模.

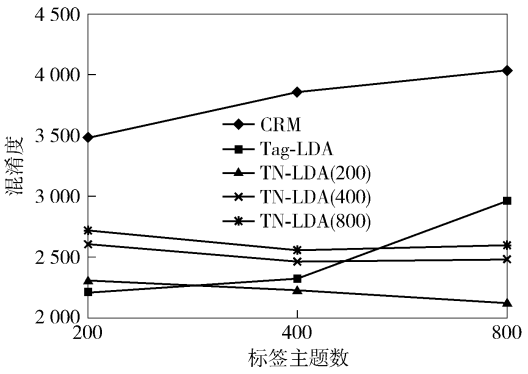

图2 BIBTEX 语料中主题和混淆度的影响

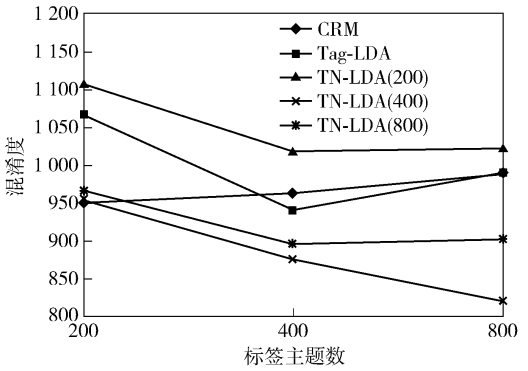

图3 Delicious 语料中主题和混淆度的影响

其次,由图3可见,一方面,与 BIBTEX 语料的情况有所不同,当词主题数为400时,TN-LDA 模型的混淆度较低,其中标签主题数为800时,混淆度仍然达到最低值,这说明,当该模型的词主题数和标签主题数分别为400和800时,对 Delicious 语料建模的能力最强. 另一方面,与 BIBTEX 数据类似,CRM 的混淆度仍是随着主题数的增加而增加的,这说明 CRM 过拟合该数据. 而 TN-LDA 模型的混淆度仍然是随着主题数的增加而下降,而且在词主题数为400和800时均低于 Tag-LDA 和 CRM 的混淆度,这

① http://www.bibsonomy.org/

② https://delicious.com/

同样说明在 Delicious 语料上,在恰当的主题数下,同时考虑了标签主题粒度和标签噪声的 TN-LDA 模型比其他模型的建模能力强. 可能的原因是 TN-LDA 模型可以在噪声标签现象较明显的 Delicious 语料上,发挥出其对噪声建模的优势.

再次,通过在 BIBTEX 和 Delicious 语料上的比较可以发现,一方面,混淆度均在标签主题数为 800 时达到最低,这印证了社会化标签多是一些表示具体概念的词,其主题粒度要小于文档中的词. 另一方面,在 BIBTEX 语料上,混淆度随着词主题数的增大而升高,而在 Delicious 语料上,词主题数取适中的 400 时达到最优,这可能的原因是: BIBTEX 的文档来源于学术出版物中的摘要,用词较规范,且频繁出现一些主题粒度较大的抽象词; 而 Delicious 语料来源为网页,领域更加自由,既有粒度较小的词,也有粒度较大的词.

2.3　标签推荐性能比较

实验中 TN-LDA 模型的词主题数选取第 1 组对比实验中混淆度最低的结果.

首先考察 BIBTEX 语料上的标签推荐性能. 由表 1 可知,随着标签主题数的增加,TN-LDA 相比 Tag-LDA 提升较显著,说明 BIBTEX 数据中标签的主题粒度较小,更倾向于描述具体概念.

表 1　BIBTEX 语料标签推荐性能

模型	K	MAP@5	@10	@15	@20
CRM	200	9.3	10.2	10.5	10.6
	400	9.9	11.2	11.4	11.6
	800	10.4	11.6	11.9	12.1
Tag-LDA	200	15.8	17.8	18.6	18.9
	400	18.5	20.8	21.6	22
	800	20.5	22.8	23.4	23.6
TN-LDA	200	18.7	20.5	21.0	21.3
	400	20.9	22.6	23.2	23.6
	800	22.6	25.1	26.6	26.9

其次考察 Delicious 语料上的标签推荐性能. 通过观察表 2,TN-LDA 相比 Tag-LDA,提升较大. 这主要由于 Delicious 语料中社会化标签的噪声现象较明显,而且采用不同粒度的主题模型增大了模型的灵活性. 同样的,随着标签主题数的增加,性能提升不是很显著,说明 Delicious 语料的标签虽然也比较倾向于较细粒度的标签,但是由于其标签间的主题粒度差异较大,提升程度不如 BIBTEX 明显.

表 2　Delicious 语料标签推荐性能

模型	K	MAP@5	@10	@15	@20
CRM	200	13.1	14.9	15.6	16.2
	400	14.0	16.3	17.0	17.8
	800	15.8	18.4	19.2	19.8
Tag-LDA	200	11.9	14.4	15.8	16.5
	400	13.9	17.2	18.5	19.3
	800	15.5	19.1	20.5	21.3
TN-LDA	200	18.4	20.6	22.1	22.3
	400	19.3	22.1	22.8	22.9
	800	20.5	22.7	23.2	23.3

3　结束语

针对社会化标签语料的特点,TN-LDA 模型分别在标签主题粒度与文档不一致,存在部分噪声标签两个方面的问题上进行了改进. 实验表明,在 BIBTEX 和 Delicious 语料上,TN-LDA 模型,相比未同时对上述两个问题建模的 CRM 和 Tag-LDA 模型,在衡量建模能力的混淆度和标签推荐的 MAP 值这两个指标上均有所提高.

参考文献:

[1] Song Yang, Zhuang Ziming, Li Huajing, et al. Real-time automatic tag recommendation [C] // Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. Singapore: ACM Press, 2008: 515-522.

[2] Si Xiance, Sun Maosong. Tag-LDA for scalable real-time tag recommendation [J]. Journal of Information & Computational Science, 2009, 6(1): 23-31.

[3] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[4] Blei D M, Jordan M I. Modeling annotated data [C] // Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto: ACM Press, 2003: 127-134.

[5] Chen Xin, Lu Caimei, An Yuan, et al. Probabilistic models for topic learning from images and captions in online biomedical literatures [C] // Proceedings of the 18th International Conference on Information and Knowledge Management (CIKM). Hong Kong: ACM Press, 2009: 495-504.

[6] Iwata T, Yamada T, Ueda N. Modeling social annotation

data with content relevance using a topic model [C] // Proceedings of the 24th Annual Conference on Neural Information Processing Systems (NIPS). Vancouver: MIT Press, 2009: 835-843.

[7] Krestel R, Fankhauser P, Nejdl W. Latent dirichlet allocation for tag recommendation [C] // Proceedings of 3rd ACM Conference on Recommender Systems (RecSys). New York: ACM Press, 2009: 61-68.

[8] Wetzker R, Zimmermann C, Bauckhage C. Analyzing social bookmarking systems: a del. icio. us cookbook [C] // Proceedings of the 18th European Conference on Artificial Intelligence (ECAI). Amsterdam: IOS Press, 2008: 26-30.

(上接第26页)

[4] Giuggioli B P, Marseguerra M, Zio E. Multi-objective optimization by genetic algorithms: application to safety systems [J]. Reliability Engineering and System Safety, 2001(72): 59-74.

[5] 罗佑新，车晓毅，杨继荣，等. 高维多目标灰色稳健优化设计及其Matlab实现 [J]. 农业机械学报，2008，39(8): 157-160.
Luo Youxin, Che Xiaoyi, Yang Jirong, et al. Grey robust optimization design of high dimension multi-objective and its achieving with matlab [J]. Transactions of the Chinese Society of Agricultural Machinery, 2008, 39(8): 157-160.

[6] Teixeir A A, Cunha A E, Clemente J J, et al. Modeling and optimization of a recombinant BHK-21 cultivation process using hybrid grey-box systems [J]. Journal of Biotechnology, 2005, 118(3): 290-303.