doi:10.3969/j.issn.1672-1292.2017.01.012

基于 mixtureLDA 的微博主题挖掘

万家华

(安徽新华学院信息工程学院,安徽 合肥 230088)

[**摘要**] 针对目前的主题挖掘只考虑主题内容的概率分布方法,本文提出一种综合考虑内容、时间等因素的微博主题挖掘模型 mixtureLDA. 该模型能够分析用户不同类型微博的主题概率分布和时间微博主题概率. 实验使用新浪微博数据集,结果表明基于 mixtureLDA 的微博主题挖掘模型能够有效地挖掘出用户微博和时间微博的主题概率分布. 与 MB-LDA、userLDA 模型对比,mixtureLDA 模型可有效降低困惑度.

[**关键词**] 微博,主题挖掘,微博类型,mixtureLDA
[**中图分类号**]TP391.6 [**文献标志码**]A [**文章编号**]1672-1292(2017)01-0080-06

Topic Model Based on MixtureLDA in Microblog Platform

Wan Jiahua

(Institute of Information Engineering, Anhui Xinhua University, Hefei 230088, China)

Abstract: Current studies on the topic model provide little discussion on time factor, but only focus on content factor. In this paper, a comprehensive content-considered and time-considered topic model, mixtureLDA is presented. Through the model one can obtain different kinds of user microblogs and time microblogs of topic probability distribution. The statistic data derived from Sina Weibo are applied as a case study. The results show that the topic model based on mixtureLDA provides more reliable topic probability distribution of user microblogs and time microblogs. Compared with MB-LDA and user LDA methods, the perplexity value from mixtureLDA method is lower, which means that it is more effective.

Key words: Microblog, topic model, microblog types, mixtureLDA

自 2006 年 Twitter 诞生以来,微博[1]已成为人们社交生活中必不可少的一项社交活动. 社交用户通过浏览器、手机、平板电脑等方式登陆微博客户端,发布自己对周围发生事件的感受. 微博内容反映了用户对最新事件的直观感受,也反应了用户的一些潜在兴趣偏好.

微博目前主要有以下 4 种类型:原创微博、转发微博、对话微博、话题微博. 原创微博一般是用户对发生在身边的事情的直观感受;转发微博更多倾向于媒体信息的推广;对话微博通过@好友的形式与好友互动交流;话题微博表达了用户对一个微博话题的观点. 微博的近社交关系,让用户认为微博上发布的信息可信度较高,并在浏览过程中渐渐影响用户的潜在行为.

微博主题[2]是对微博内容的概括,微博主题能够反映用户关注的一系列微博内容的核心思想. 因此,如何准确高效地挖掘出用户微博的主题[3]已成为学术界和工业界想要解决的一个热点话题. 目前针对微博主题的研究主要在于研究爆炸性激增主题和平稳主题检测以及利用微博主题进行社区群体的挖掘. 这些研究大都只是微博与用户绑定求解主题概率,而未充分考虑到用户发布微博的类型与用户主题之间的关系.

本文在研究 LDA 模型[4]的基础上,结合新浪用户微博内容类别特征与时间特征,提出一种 mixtureLDA 微博主题挖掘模型,在分析用户微博内容特征的同时结合时间特征,为不同微博类型特征设定不同的超参数,使得模型挖掘的微博主题更贴近用户的真实兴趣倾向. 本文数据采集自新浪微博开发平台,该模型在真实数据集上取得了良好的实验结果,能够分析出用户的微博主题和共同关注的热点话题,可将结果应用于微博主题的个性化推荐.

收稿日期:2016-08-08.
基金项目:安徽省高校自然科学重点项目(KJ2014A100).
通讯联系人:万家华,讲师,研究方向:数据挖掘、Web 信息处理. E-mail:349826355@qq.com

1 相关工作

随着微博的普及使用,微博的社交价值受到国内外学者的广泛关注. 主题模型[5](topic model)在自然语言处理领域是用来在文本集合中发现潜在主题的一种概率图统计模型. 针对主题模型的理论研究主要是以 Blei 的 LDA[6]最为突出,Blei 在 2003 年提出 LDA 模型让主题模型受到关注. LDA 模型采用“词袋”概念,将每篇文档视为一个主题向量,从而将文本信息转化为数值信息便于后续的工作研究. Blei 在随后的几年里一直致力于推进 LDA 模型的理论优化以适应不同场景应用的需求,如 CTM[7](correlated topic models)、DTM[8](dynamic topic models)、online LDA[9]等.

由于主题模型能够分析出给定文档集合的潜在主题,因此被广泛应用于文本主题挖掘[10]、文本分析[11]、社交网络分析[12]、文本分类[13]等领域. 微博作为社交网络的典型应用,针对微博内容的研究一直是学者们关注的热门话题. Paul[14]开发了 Who Gives a Tweet 网站收集用户来自粉丝和游客的评价,根据评价内容进行统计分析,从而了解微博内容的真正价值. 针对微博内容的研究也在不断深化,Daniel 利用 Labeled LDA[15]分析精神、风格、状态、社交 4 种类别的微博主题. 张晨逸将转发、原创、对话微博的特征加入考虑,设计了 MB-LDA[16]模型来分析各特征下的微博主题. Wayne 使用 TwitterLDA[17]分析 Twitter 的主题,并对比分析 Twitter 与传统媒体数据的区别.

另有一些学者关注了微博内容主题与时间的关系,分析微博内容的平稳主题、共享主题和爆炸性主题. 如 TimeUserLDA[18]模型分析了用户微博的主题分布及随时间变化的爆炸性主题分布,Unified Model[19]则利用临时信息分析出社交媒体的平稳主题和爆炸性主题. 两个模型都主要针对爆炸性话题识别,在模型求解方法上分别使用 gibbs 抽样和 EM 算法求解模型参数.

与本文有关的模型为 MB-LDA、TimeUserLDA、Unified Model. MB-LDA 模型分析了转发、原创、对话微博类别的主题概率分布,但未将微博与用户关联,无法直接分析用户微博主题概率分布[20]. TimeUserLDA 和 Unified Model 虽然分析了用户主题概率分布和热点话题主题概率分布,却未考虑用户微博内容类别特征. 本文模型在考虑用户微博内容类别特征的基础上,综合考虑时间对用户微博主题的影响,提出一种基于 mixtureLDA 的微博主题挖掘模型. 通过 mixtureLDA 模型能够分析用户在时间和微博内容特征共同影响下的主题概率分布,获知用户关注的话题和关键词汇.

2 基于 mixtureLDA 的微博主题挖掘模型

2.1 模型定义

本文将新浪用户微博内容类别特征主要划分为 4 类:普通、对话、转发、话题微博. 同时考虑到用户主题会受时间因素变化,加入了时间主题概率分布分析. 与模型有关的定义如下:

定义 1 微博集合 M 中任一微博 m_j 由词语集合 V 中的单词以词袋模型组合而成,即 $m_j=\{v_1,v_2,\cdots,v_{|m|}\}$,其中 $v_m\in V(1\leqslant w\leqslant|m|)$,$m_j\in M$.

定义 2 给定一个用户集合 U,每个用户 u_i 可发布上述 4 种类型的微博内容,M_{n}、M_{c}、M_{r}、M_{tp}分别代表普通、对话、转发、话题微博. 通过 mixtureLDA 模型将会得到对应一组微博内容类型有关的主题概率分布 θ_{n}、θ_{c}、θ_{r}、θ_{tp}.

定义 3 在给定的时间间隔 $1,\cdots,T$ 内,用户 u_i 在 t_k 时刻发表微博 m_j,其中 $1\leqslant t_k\leqslant T$. 在时间间隔 1~$T$ 内,所有用户发表的微博为时间微博类型,代表所有用户普遍关注的主题,δ^{t} 代表时间微博的主题概率分布.

mixtureLDA 模型是在 LDA 理论和分析新浪微博内容特征的基础上,分析用户在各微博内容类别特征下的主题概率分布,以及全体用户在一段时间内的主题概率分布. 通过该模型分析用户在各微博内容特征下的主题概率和用户关注的主题词汇,为用户微博的个性化推荐提供理论依据.

2.2 mixtureLDA 主题模型

在 mixtureLDA 模型中定义了 5 种主题类型,分别为普通、对话、转发、话题、时间主题. 普通、对话、转发、话题主题是与用户关联,用于分析用户在各微博内容类型下对应的主题概率分布. 时间主题用于分析全体微博用户在一段时间内关注的普通话题,了解用户普遍关注的主题. mixtureLDA 贝叶斯网络图模型

结构如图 1 所示,没有加入时间因素的微博主题挖掘贝叶斯图模型如图 2 所示.

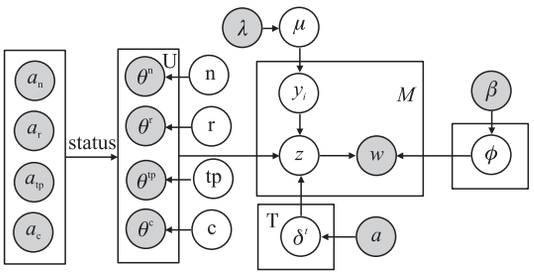

图 1　mixtureLDA 贝叶斯图模型

Fig. 1　mixtureLDA Bayesian graph model

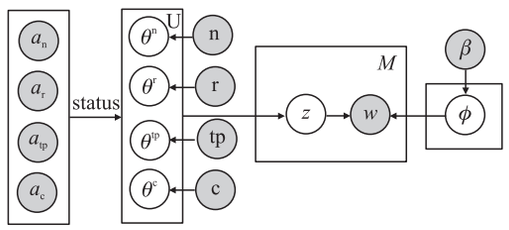

图 2　userLDA 贝叶斯图模型

Fig. 2　userLDA Bayesian graph model

所有微博都有 C 个潜在主题,每个主题 c 对应一个词概率分布 ϕ^c. 微博的生成过程是首先从参数为 π 的伯努利分布中抽样,确定当前微博是时间微博还是用户微博,$y_i=0$ 为用户微博,$y_i=1$ 为时间微博. 若当前微博为时间微博,则服从 δ^t 多项式分布. 微博生成的单词按 δ^t 主题分布从词概率分布 ϕ 中获取单词,生成当前微博. 否则,判断当前用户微博为哪种微博内容类型. 若 $s_i=1$,则当前微博为普通微博,则按 θ_n 主题分布从词概率分布 ϕ 中获取单词生成微博. 对应的对话、转发、话题微博的生成步骤与普通微博的生成步骤类似. 其中,status 代表用户微博内容类型,n、r、c、tp 表示用户发布普通、转发、对话、主题微博,所有微博的生成步骤流程如表 1 所示

表 1　微博生成步骤流程

Table 1　Micro-blog generated step flow

(1) for each topic $c=1,2,\cdots,C$	if $s_i=3$
draw $\phi\sim\mathrm{Dir}(\beta)$	draw $\theta^u=\theta^r\sim\mathrm{Dir}(\alpha^r)$
(2) draw $\pi\sim\mathrm{Beta}(\gamma)$	if $s_i=4$
(3) for each time point $t=1,2,\cdots,T$	draw $\theta^u=\theta^{tp}\sim\mathrm{Dir}(\alpha^{tp})$
draw $\delta^t\sim\mathrm{Dir}(\alpha)$	(5) for each microblog $i=1,\cdots,M$
(4) for each user $u=1,2,\cdots,U$	draw $y_i\sim\mathrm{Bernoulli}(\pi)$
if $s_i=1$	if $y_i=0$ $z_i\sim\mathrm{Multi}(\theta^u)$
draw $\theta^u=\theta^n\sim\mathrm{Dir}(\alpha^n)$	if $y_i=1$ $z_i\sim\mathrm{Multi}(\delta^t)$
if $s_i=2$	(6) for each word $j=1,2,\cdots,V$
draw $\theta^u=\theta^c\sim\mathrm{Dir}(\alpha^c)$	draw $w_{ij}\sim\mathrm{Multi}(\phi^{z_i})$

2.3　mixtureLDA 模型参数求解

本文采用 gibbs[21] 抽样方法求解 mixtureLDA 模型隐含参数,即抽样主题迭代计算,最终得到模型求解的各主题概率分布和词概率分布结果. mixtureLDA 模型主要的参数推导过程如下:

首先,对于一条微博,通过一个伯努利分布来随机分配当前微博类别为用户微博或时间微博. 引入隐含变量 y_i 区分用户和时间微博,1 为时间微博,0 为用户微博. 采样时间微博主题时,计算在当前主题下时间微博在非当前主题各隐含变量的后验概率,通过式(1)抽样迭代获得时间微博的主题概率:

$$p(y_i=1,z_i=c\mid z_{\neg i},y_{\neg i},w)\propto\frac{M^{\mu}_{(y_i)}+\lambda}{M^{\mu}_{(.)}+2\lambda}\cdot\frac{M^{m}_{(c)}+\alpha}{M^{m}_{(.)}+C\alpha}\cdot\frac{M^{c}_{(v)}+\beta-1}{M^{c}_{(.)}+V\beta-1},\tag{1}$$

其中,$M^u_{\cdot}$ 代表用户发表的微博总数,M^u_1 代表用户发表时间微博总数,M^u_0 代表用户发表时间微博总数. $M^m_{\cdot}$ 代表时间微博总数,M^m_c 代表在 c 主题下时间微博个数. $M^c_{\cdot}$ 代表 c 主题下单词总数,M^c_v 代表单词在 c 主题中出现的次数,V 代表微博单词总数.

其次,模型中用户微博分 4 类,引入隐含变量 s_i 来区分 4 种不同类型的微博. $s_i=1$ 为普通微博,$s_i=2$ 为对话微博,$s_i=3$ 为转发微博,$s_i=4$ 为主题微博. 抽样主题流程与时间微博主题抽样流程类似,微博主题概率如式(2)所示:

$$p(y_i=0,s_i=x,z_i=c\mid z_{\neg i},s_{\neg i},y_{\neg i},w)\propto\frac{M^{\mu}_{(y_i)}+\lambda}{M^{\mu}_{(.)}+2\lambda}\cdot\frac{M^{(y_i)}_{(x)}+\gamma}{M^{(y_i)}_{(.)}+4\gamma}\cdot\frac{M^{m}_{(c)}+\alpha}{M^{m}_{(.)}+C\alpha}\cdot\frac{M^{c}_{(v)}+\beta-1}{M^{c}_{(.)}+V\beta-1}.\tag{2}$$

其中,λ 的引入是为了平衡时间微博和用户微博的比例,γ 的用途在于平衡用户微博类型的比例. $M^m_{\cdot}$ 代表

用户微博总数,M_c^m 代表在 c 主题下用户微博的总数.

最终,模型通过 gibbs 抽样方法的迭代计算,得到各类型的主题概率分布和词概率分布结果,可得 ϕ、θ_n、θ_c、θ_r、θ_{tp}、δ_t 的具体计算公式为:

$$\varphi_z \propto \frac{M_v^c+\beta-1}{M_\cdot^c+V\beta-1}, \tag{3}$$

$$\theta_n = \frac{M_n^c+\alpha_n}{M_\cdot^c+C\alpha_n}, \tag{4}$$

$$\theta_c = \frac{M_{c0}^c+\alpha_c}{M_\cdot^c+C\alpha_c}, \tag{5}$$

$$\theta_r = \frac{M_r^c+\alpha_r}{M_\cdot^c+C\alpha_r}, \tag{6}$$

$$\theta_{tp} = \frac{M_{tp}^c+\alpha_{tp}}{M_\cdot^c+C\alpha_{tp}}, \tag{7}$$

$$\delta_t = \frac{M_t^c+\alpha}{M_\cdot^c+C\alpha}. \tag{8}$$

通过 mixtureLDA 最终获得时间微博和用户微博的主题概率分布,从而可分析用户普遍关注的主题和用户在指定微博类型下的主题概率.

3 实验

3.1 预处理

本文实验通过新浪微博 API 接口抽取新浪微博数据. 数据集包含了 170 多万用户信息和 200 万的微博信息.

为了使数据更规范便于后续模型分析,本文做如下预处理操作:

(1)去除噪声信息. 微博内容分词在微博内容中存在一些噪声数据如链接、表情、@用户,本文实验剔除了这些信息.

(2)微博内容分词. 实验使用 IKAnalyzer 对微博内容进行分词,同时加载中英文停用词典去除一些出现频率高但对主题挖掘无贡献的词语.

(3)标注微博类别. 根据微博字段的转发字段值区分当前微博是否为转发微博,1 为是,0 为否,并将该微博的类别标识为 1. 在区分转发和非转发微博后,根据微博内容是否包含#**#和@符号来标注话题微博和对话微博,分别用 2 和 3 标注,不包含这些特征符号的则为普通微博,标注为 0.

3.2 实验结果

本文模型主题数遵循经验值,取 $C=50/\alpha$,时间间隔 $T=12$. 通过训练集结果最终确定超参数为 $\alpha=\alpha^n=\alpha^c=\alpha^r=\alpha^{tp}=5$, $\lambda=5$, $\gamma=5$,确定主题个数 C 为 10. 模型评价指标采用困惑度(Perplexity)指标来评估模型,取值越小,模型推广性越好. 困惑度计算公式为:

$$\text{Perplexity}(w)=\exp\left(-\frac{\sum_m \ln P(w_m)}{\sum N_m}\right), \tag{9}$$

其中,W 是测试集,w_m 是测试集中各微博类别中的单词,N_m 为测试集各微博类别微博中的单词总数. 训练集中主题数与困惑度关系的实验结果如图 3 所示,从图中确定主题个数 C 为 10 最佳,测试集中迭代次数为 150 次最佳.

mixtureLDA 模型的主题挖掘模型效果如图 4 所示,共挖掘 10 个主题. 图 5 展示一个简单的 mixtureLDA 模型处理实例. 根据各个主题对应的高频词汇可以发现各主题的主旨,Topic 1 与微博有关,Topic3 与新年有关,该主题下的高频词汇新年、红包对应了这个主题. 通过 mixtureLDA 模型能够得到用户在普通、对话、转发、话题微博下的主题概率分布以及时间微博的概率分布.

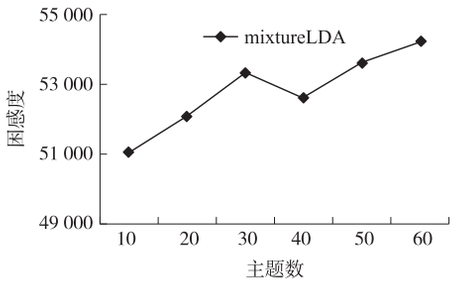

图 3　训练集主题数与困惑度关系

Fig. 3　The relationship between the number of training sets and perplexity

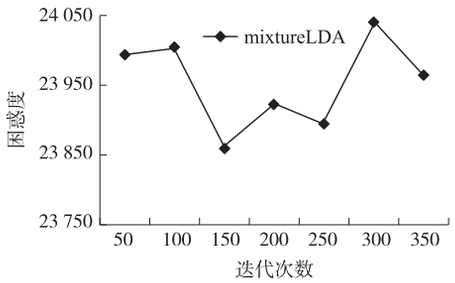

图 4　测试集主题数与迭代次数关系

Fig. 4　Relationship between the number of test sets and the number of iterations

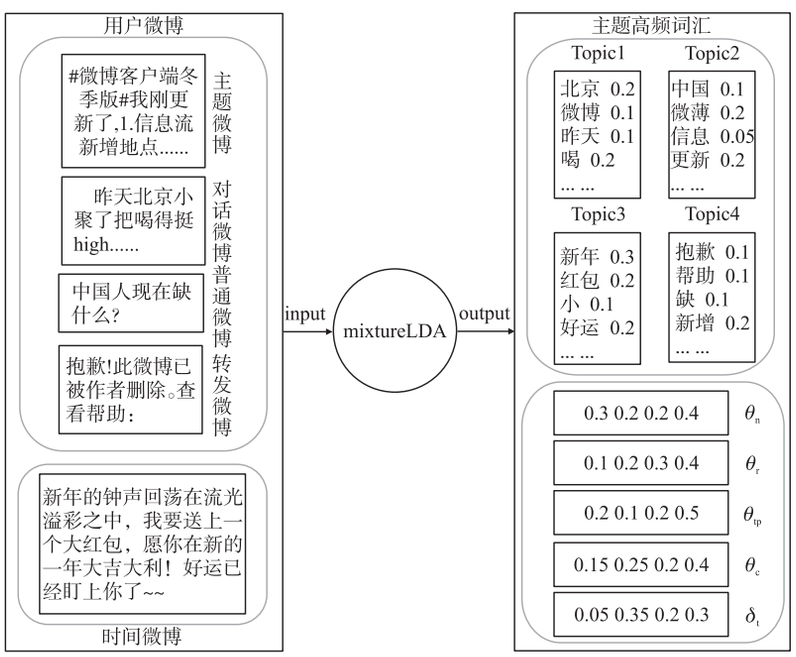

图 5　mixtureLDA 模型整体效果

Fig. 5　The overall effect of model mixtureLDA

3.3　对比实验

本文在相同模型参数设置下,对比分析 mixtureLDA、MB-LDA、userLDA 模型的效果,以困惑度作为度量标准. 模型的对比实验结果和折线图如表 2 和图 6 所示.

表 2　对比实验结果

Table 2　Comparative experimental results

迭代次数	MB-LDA	userLDA	mixtureLDA
50	29 670.442 3	24 505.700 1	23 992.471 4
100	29 670.438 6	24 544.874 5	24 003.452 6
150	29 670.212 9	24 525.312 4	23 858.175 3
200	29 670.368 1	24 530.987 0	23 922.058 8
250	29 670.388 3	24 488.979 5	23 893.365 3
300	29 670.289 2	24 541.082 2	24 040.248 1
350	29 673.897 1	24 522.467 2	23 963.397 0

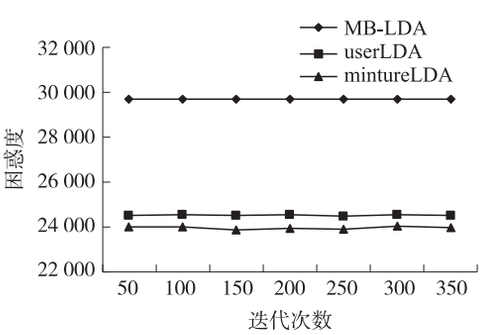

图 6　模型困惑度对比折线图

Fig. 6　Comparison of model confusion

从表 2 和图 6 可以看出,mixtureLDA 模型比 MB-LDA 和 userLDA 模型的困惑度值低,都在迭代 150 次达到最佳值. mixtureLDA 的整体收敛效果很平稳,能够有效地挖掘用户微博和时间微博的主题,易于推广该模型进行主题挖掘应用.

如图 6 所示,mixtureLDA 模型的困惑度值最低,能够以较低的困惑度值快速收敛至最佳值,并分析出用户在各微博类型下的主题概率和时间微博概率,在基于主题的个性化推荐中有很强的应用性.

4 结语

本文针对新浪微博内容特点,将微博划分为用户微博和时间微博. 然后根据用户微博的类型特点细分为普通、转发、对话、话题微博,通过 mixtureLDA 模型分析了全体用户关注的时间微博主题和用户在各微博类型下的主题概率. 根据主题概率分析的高频词汇可为基于主题的微博个性化推荐提供依据.

今后的研究工作中将继续优化 mixtureLDA 模型的效率,同时考虑微博内容表情和短链对主题的影响,将 mixtureLDA 主题分析的结果应用于个性化推荐领域.

[参考文献](References)

[1] GUO Z,LI Z,TU H. Sina microblog:an information-driven online social network[C]//2011 International Conference on Cyber-worlds(CW). Calgary,Canada,2011:160-167.

[2] SHEN Y,LI S,ZHENG L,et al. Emotion mining research on microblog[C]//1st IEEE Symposium on Web Society. Lanzhou,China:IEEE,2009:71-75.

[3] NALLAPATI R,COHEN W W. Link-plsa-lda:a new unsupervised model for topics and influence of blogs[C]//Proceedings of ICWSM. Washington DC,USA,2008:84-92.

[4] 王永贵,张旭,任俊阳,等. 结合微博关注特性的 UF_AT 模型用户兴趣挖掘研究[J]. 计算机应用研究,2015,32(7):1 982-1 985.
WANG Y G,ZHANG X,REN J Y,et al. Research on micro-blog user's interest mining based on UF_AT model which combining with focusing feature of microblog[J]. Application research of computers,2015,32(7):1 982-1 985.(in Chinese)

[5] STEYVERS M,GRIFFITHS T. Probabilistic topic models[J]. Handbook of latent semantic analysis,2007,427(7):424-440.

[6] BLEI D M,NG A Y,JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research,2003(3):993-1 022.

[7] LAFFERTY J D,BLEI D M. Correlated topic models[J]. Advances in neural information processing systems,2005,18:147-154.

[8] DAVID M B,JOHN D. Lafferty:dynamic topic models[C]//Proceedings of ICML 2006. Pittsburgh,USA,2006:113-120.

[9] MATTHEW D H,DAVID M B,FRANCIS R B. Online learning for latent dirichlet allocation[C]//Proceedings of NIPS 2010. Vancouver,Canada,2010:856-864.

[10] PAK A,PAROUBEK P. Twitter as a corpus for sentiment analysis and opinionmining[C]//Proceedings of LREC 2010. Valletta,Malta,2010:1 320-1 326.

[11] LIU Z,YU W,CHEN W,et al. Short text feature selection for microblog mining[C]//2010 International Conference on Computational Intelligence and Software Engineering. Wuhan,China,2010:1-4.

[12] ZHANG Y,WU Y,YANG Q. Community discovery in twitter based on userinterests[J]. Journal of computational information systems,2012,8(3):991-1 000.

[13] LI W,SUN L,FENG Y,et al. Smoothing lda model for text categorization[C]//Proceedings of AIRS 2008. Harbin,China,2008:83-94.

[14] ANDRÉ P,BERNSTEIN M,LUTHER K. Who gives a tweet? evaluating microblogcontent value[C]//Proceedings of CSCW 2012. New York,USA,2012:471-474.

[15] RAMAGE D,DUMAIS S T,LIEBLING D J. Characterizing microblogs with topic models[C]//Proceedings of ICWSM 2010. Washington DC,USA,2010:130-137.

[16] ZHANG C,SUN J. Large scale microblog mining using distributed mb-lda[C]//Proceedings of IW3C2 2012. Lyon,France,2012:1 035-1 042.

[17] ZHAO W X,JIANG J,WENG J S,et al. Comparing twitter and traditional media using topic models[C]//Proceedings of ECIR 2011. Dublin,Ireland,2011:338-349.

[18] DIAO Q M,JIANG J,ZHU F D,et al. Finding bursty topics from microblogs[C]//Proceedings of ACL 2012. Jeju,Korea,2012:536-544.

[19] YIN H,CUI B,LU H,et al. A unified model for stable and temporal topic detection from social media data[C]//2013 IEEE 29th International Conference on Data Engineering(ICDE). Brisbane,Australia,2013:661-672.

[20] 陶永才,何宗真,石磊,等. 基于加权动态兴趣度的微博个性化推荐[J]. 计算机应用,2014,34(12):3 491-3 496.
TAO Y C,HE Z Z,SHI L,et al. Personalized microblogging recommendation based on weighted dynamic degree of interest[J]. Journal of computer applications,2014,34(12):3 491-3 496.(in Chinese)

[21] GRIFFITHS T. Gibbs sampling in the generative model of latent dirichlet allocation[R]. Palo Alto:Standford University,2002.

[责任编辑:严海琳]