计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
48(10): 1795-1802, 2011

基于 MB-LDA 模型的微博主题挖掘

张晨逸[1] 孙建伶[1] 丁轶群[2]
[1](浙江大学计算机科学与技术学院 杭州 310027)
[2](浙江大学工业技术研究院 杭州 310027)
(zhangchenyi.zju@gmail.com)

Topic Mining for Microblog Based on MB-LDA Model

Zhang Chenyi[1], Sun Jianling[1] and Ding Yiqun[2]
[1](*College of Computer Science and Technology, Zhejiang University, Hangzhou* 310027)
[2](*Industrial Technology Research Institute of Zhejiang University, Hangzhou* 310027)

Abstract As microblog grows more popular, services like Twitter have become information providers on a web scale. Early work on microblog focused more on its user relationship and community structure, without considering the value of content. So the research on microblog requires a change from solely user's relationship analysis to its content mining. Although traditional text mining methods have been studied well, no algorithm is designed specially for microblog data, which contain structured information on social network besides plain text. In this paper, we propose a novel probabilistic generative model based on LDA, called MB-LDA, which is suitable to model the microblog data and takes both contact relation and document relation into consideration to help topic mining in microblog. We present a Gibbs sampling implementation for inference of our model, and find not only the topics of microblog, but also the topics focused by contactors according to the final results. Besides, our model can be extended to many texts associated with social networking such as E-mails and forum posts. Experimental results on actual dataset show that MB-LDA model can offer an effective solution to topic mining for microblog.

Key words microblog; topic mining; LDA; probabilistic generative model; social network

摘 要 随着微博的日趋流行,Twitter 等微博网站已成为海量信息的发布体,对微博的研究也需要从单一的用户关系分析向微博本身内容的挖掘进行转变.在数据挖掘领域,尽管传统文本的主题挖掘已经得到了广泛的研究,但对于微博这种特殊的文本,因其本身带有一些结构化的社会网络方面的信息,传统的文本挖掘算法不能很好地对它进行建模.提出了一个基于 LDA 的微博生成模型 MB-LDA,综合考虑了微博的联系人关联关系和文本关联关系,来辅助进行微博的主题挖掘.采用吉布斯抽样法对模型进行推导,不仅能挖掘出微博的主题,还能挖掘出联系人关注的主题.此外,模型还能推广到许多带有社交网络性质的文本中.在真实数据集上的实验表明,MB-LDA 模型能有效地对微博进行主题挖掘.

关键词 微博;主题挖掘;LDA;概率生成模型;社交网络

中图法分类号 TP181

收稿日期:2011-06-23;**修回日期**:2011-08-23
基金项目:"核高基"国家科技重大专项基金项目(2010ZX01042-002-003)

微博作为 Web2.0 时代兴起的一种互联网社交网络服务,以其快速便捷的特性风靡全球.微博基于用户之间的关联关系,构筑了一个信息传播和分享的平台,用户可以通过网络、手机或是其他客户端登录微博,实时地进行短文本信息的更新和分享.著名的微博网站 Twitter 注册用户已达 1.75 亿,每天发布的 Twitter 消息(tweet)超过 1.3 亿条.

用户通过微博网站构建的平台可以发布自己最新的状态、表达自己对事物的观点,也可以对某人单独地发起对话,还可以转发别人的微博.按照以上不同的消息发布方式,微博主要分为 3 类[1]:广播(broadcast)、对话(conversion)和锐推(retweet).广播类型的微博最常见,所有人均可见;对话类型的微博有特定的发送对象;而锐推类型的微博则是对感兴趣微博的转发.一条普通的锐推微博如下:"LOL RT@Ethan This is a good website http://www.vlis.zju.edu.cn",其中 RT 表示锐推类型,RT 之前的是原创内容,RT 之后的是转发内容,@Ethan 表示转发部分的作者为 Ethan.

在信息爆炸时代,从海量信息中挖掘出有效的主题信息,分析出内在语义关联显得尤为重要.微博本身是一种非结构化的文本信息载体,却又带有一些结构化的社会网络方面的信息,这种社会网络的关联关系在主题挖掘时可以起到辅助作用;另一方面,每条微博可认为是一个文本片段(通常只有一句话),携带的信息量不大,这种短文本结构会加大其主题挖掘的难度[1].以上这些特性决定了微博主题挖掘不能简单地套用传统的文本主题挖掘的方法.以上文中那条微博为例,单单凭一个单词 LOL(laugh out loud)并不能有效地挖掘出微博原创内容的主题信息,但是如果意识到它是一条锐推类型的微博,如果把转发内容综合考虑进来,就能推断出原创部分的主题与网站有关.

目前对于微博的研究也大多停留在用户关系和社区结构的分析上,很少有针对微博发布内容的研究[2],很大程度上是因为传统的文本挖掘算法多用于传统的语料库,不考虑微博文本内蕴含的特殊的结构化信息,不能很好地对微博数据进行建模.本文在研究 LDA(latent Dirichlet allocation)的基础上,结合微博的以上特性,综合考虑了微博的联系人关联关系和文本关联关系(见 3.2 节定义 1 和定义 2),提出了一种挖掘微博主题的新模型 MB-LDA (MicroBlog-latent Dirichlet allocation).

本文的贡献如下:

1) 综合考虑了微博中的结构化数据(联系人信息和锐推信息)和非结构化数据(文本信息),提出了适合于微博主题挖掘的新模型 MB-LDA;

2) 利用吉布斯抽样法(Gibbs sampling)对 MB-LDA 模型进行求解,实现主题挖掘,并可以将模型推广到其他带有社会网络性质的文本中(如 Email、聊天记录等);

3) 在真实数据集上对模型进行了验证,表明 MB-LDA 模型能很好地对微博数据进行主题挖掘.

1 相关工作

近年来,关于文本主题挖掘的方法受到了人们广泛的关注和研究,各类算法不断涌现.

1.1 传统的主题挖掘算法

传统的主题挖掘最早可以追溯到采用文本聚类的算法,通过 VSM(vector space model)将文本里的非结构化数据映射到向量空间中的点,然后用传统的聚类算法实现文本聚类.文本聚类有基于划分的算法(如 K-means 算法)、基于层次的算法(自顶向下和自底向上算法)、基于密度的算法等等[3].聚类结果可以近似认为满足同一个主题.但是,这种基于聚类是算法普遍依赖于文本之间距离的计算,而这种距离在海量文本中是很难定义的;此外,聚类结果也只是起到区分类别的作用,并没有给出语义上的信息,不利于人们的理解.

1.2 基于线性代数的主题挖掘算法

LSA(latent semantic analysis)是 Deerwester 等人[4]提出的一种基于线性代数挖掘文本主题的新方法.LSA 利用 SVD(singular value decomposition)的降维方法来挖掘文档的潜在结构(语义结构),在低维的语义空间里进行查询和相关性分析,通过 SVD 等数学手段,使得这种隐含的相关性能够被很好地挖掘出来.研究显示[5],当这个语义空间的维度和人类语义理解的维度相近时,LSA 能够更好地近似于人类的理解关系,也就是说,将表面信息转化为深层次的抽象.

LSA 的局限性在于它不能解决文本的"一词多义"问题,因为一个单词在语义空间中只有一个坐标(可以认为是该单词多个意义的平均),无法用多个坐标来表示多个意义;且 SVD 涉及到矩阵运算,计算开销较大,且计算结果在很多维度上为负数,使得主题的理解并不直观.

1.3 基于概率模型的主题挖掘算法

主题模型(topic model)是一种使用概率的产生式模型来挖掘文本主题的新方法[6]. Topic Model 中假设,主题可以根据一定的规则生成单词,那么在已经知道文本单词的情况下,可以通过概率方法反推出文本集的主题分布情况. 最具代表性的 Topic Model 是 PLSA 和 LDA.

PLSA (probabilistic latent semantic analysis)是 Hofmann[7]在研究 LSA 的基础上提出的基于最大似然法(maximun likelihood)和产生式模型(generative model)的概率模型. PLSA 沿用了 LSA 的降维思想:在常用的文本表达方式(tf. idf)[8]下,文本是一种高维数据;主题的数量是有限的,对应低维的语义空间,主题挖掘就是通过"降维"将文档从高维空间投影到了语义空间. PLSA 通常运用 EM 算法对模型进行求解. 在实际运用中,由于 EM 算法的计算复杂度小于传统 SVD 算法,PLSA 在性能上、在处理大规模数据方面也通常优于 LSA.

LDA 在 PLSA 的基础上加入了 Dirichlet 先验分布,是 PLSA 的一个突破性的延伸. LDA 的创始者 Blei 等人[9]指出,PLSA 在文档对应主题的概率计算上没有使用统一的概率模型,过多的参数会导致过拟合(overfitting)现象,并且很难对训练集以外的文档分配概率. 基于这些缺陷,LDA 引入了超参数,形成了一个"文档-主题-单词"3 层的贝叶斯模型,然后通过运用概率方法对模型进行推导,来寻找文本集的语义结构,挖掘文本的主题.

Topic Model 被广泛地应用于主题挖掘[9]、文本检索[10]、文本分类[9]、引文分析[11]和社交网络分析[12]等领域,此外还应用于处理非文本信息,包括计算机视觉、图像等[13]领域.

近年来对 Topic Model 的研究也不断深化,衍生出了各式各样的模型,如 Dynamic Topic Model[14], Syntactic Topic Nodel[15]等等. 与本文工作比较相关的是考虑了文本之间关系的两个模型:Link-PLSA-LDA 和 HTM(hypertext topic model).

Link-PLSA-LDA 是 Nallapati 和 Cohn[16]提出的用于引文分析的 Topic Model. 在此模型中,被引用文本是用 PLSA 生成的,引用文本是用 LDA 生成的,且模型假设两者享有相同的主题.

HTM 是由 Sun[17]等人提出的适用于超文本分析的 Topic Model. HTM 在生成文本的过程中加入了超链接的影响因素,能够更好地对超文本进行主题挖掘和文本分类.

Table 1 Notation

表 1 符号定义说明

Parameter	Definition
α,α_c	Hyperparameters for θ_d and θ_c
β	Hyperparameter for φ
T	Number of topics
D	Number of document/microblog
V	Number of word
c	Contactor in conversation messages(@)
λ	Weight parameters for retweet messages
θ_c	Topic distribution associated with contactor c
θ_d	Topic distribution over doc/microblog d
$\theta_{d_{\mathrm{RT}}}$	Topic distribution over retweet message d
φ	Word distribution over topics
$z_i\ z_{-i}$	Topic of word i (indicators before sampling i)
w	Word in documents/microblogs
$n_{d,j}(n_{d,\cdot})$	Co-occurrence of d and topic j(all topics)
$n_{c,j}(n_{c,\cdot})$	Co-occurrence of c and topic j(all topics)
$n_{j,v}(n_{j,\cdot})$	Co-occurrence of j and word v(all words)
$\pi_c(r)$	Bool parameters to decide specific microblogs

2 微博主题挖掘

2.1 文本生成模型 LDA

主题模型采用概率的产生式模型来对文本进行建模,它的基本思想是[6]每个文本都可以表示成一系列主题的混合分布,记为 $P(z)$;同时每个主题是词汇表中所有单词上的概率分布,记为 $P(w|z)$. 因此,一个文本中每个单词的概率分布如式(1)所示:

$$P(w_i)=\sum_{j=1}^{T}P(w_i\mid z_i=j)P(z_i=j). \quad (1)$$

LDA 是第 1 个完备的主题模型,它生成文本的方式可以用图 1 中的贝叶斯网络图来表示. 最开始,LDA 从参数为 β 的 Dirichlet 分布中抽取主题与单词的关系 φ. LDA 生成一个文本时,首先从参数为 α 的 Dirichlet 分布中抽样出该文本 d 与各个主题之间的关系 $\boldsymbol{\theta}_d$,当有 T 个主题时,$\boldsymbol{\theta}_d$ 是一个 T 维向量,每个元素代表主题在文本中的出现概率,满足 $\sum_T\theta_{d_T}=1$;接着,从参数为 $\boldsymbol{\theta}_d$ 的多项式分布中抽

样出当前单词所属的主题 z_{dn}；最后从参数为 $\varphi_{z_{dn}}$ 的多项式分布中抽取出具体单词 w_{dn}. 一个文本中所有单词与其所属主题的联合概率分布如式(2)所示：

$$P(w,z\mid\alpha,\beta)=P(w\mid z,\beta)P(z\mid\alpha)\int P(z\mid\theta)P(\theta\mid\alpha)\mathrm{d}\theta\int P(w\mid z,\varphi)P(\varphi\mid\beta)\mathrm{d}\varphi. \tag{2}$$

如图 1 所示，在 LDA 中，文本的单词是可观测到的数据，而文本的主题是隐式变量. 根据文本的生成规则和已知数据，LDA 通过概率推导可以求得文本的主题结构. 常用的推导方法有变分贝叶斯(variational Bayesian)、吉布斯抽样(Gibbs sampling)、期望值传播(expectation propagation)等等[9,18-19].

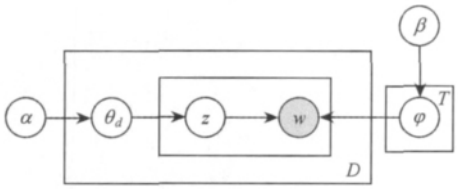

Fig. 1 Bayesian network of LDA.

图 1 LDA 模型的贝叶斯网络图

2.2 微博生成模型 MB-LDA

微博不同于一般的文本，本身带有表征微博之间关联的文字信息：@和 RT. @表示微博之间的联系人关联关系，RT 则表示微博之间的文本关联关系，两种关联关系分别定义如下：

定义 1. 微博的联系人关联关系指的是带有@的微博与@的联系人之间存在潜在的语义关联. 一般来说，与同一个联系人存在关联的微博，他们的主题往往也是相关的. 这种关联关系常见于对话类型的微博. 如两条对话微博："@Ethan Can you lend me a book on data mining"和"@Ethan HELP me on these computer exercises"，如果考虑了联系人关联关系，就能把两条看似无关的微博联系在一起，推断出第 2 条微博中的 computer exercises 与数据挖掘有关.

定义 2. 微博的文本关联关系指的是带有 RT 的微博与原微博之间存在潜在的语义关联. 一般来说，转发部分和原创部分的主题往往是相关的. 这种关联关系常见于锐推类型的微博. 如一条锐推微博："Good job RT@Ethan I have finished this experiment"，对原创部分的内容"Good job"很难有效地挖掘主题，但通过文本关联关系，联系转发部分的内容，可以推断原创部分的 job 是一项实验工作.

MB-LDA 是在研究 LDA 的基础上，对微博的联系人关联关系和文本关联关系进行统一建模，形成的适合于微博主题挖掘的模型，模型的参数如表 1 所示. 它的贝叶斯网络图如图 2 所示. 其中 c 和 r 分别用来表征联系人关系和转发关系. 最开始，MB-LDA 从参数为 β 的 Dirichlet 分布中抽取主题与单词的关系 φ. MB-LDA 生成一条微博时，首先根据@来判断联系人关联关系：如果微博以@开头(由于对话微博多以@开头，且@在其他位置时难以判断是否表达为对话关系，因此本文只考虑以@开头的联系人关联关系)，π_c 取 1，表示为一条对话微博，就从参数为 α_c 的 Dirichlet 分布中抽样出该联系人 c 与各个主题之间的关系 θ_c，且把它赋值给微博 d 与各个主题之间的关系 θ_d；否则 π_c 取 0，直接从参数为 α 的 Dirichlet 分布中抽样出该微博 d 与各个主题之间的关系 θ_d.

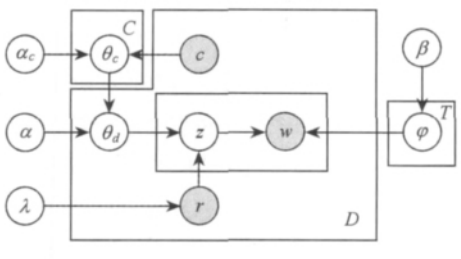

Fig. 2 Bayesian network of MB-LDA.

图 2 MB-LDA 模型的贝叶斯网络图

整个微博集中，θ 的概率分布如式(3)所示：

$$P(\theta\mid\alpha,\alpha_c,c)=P(\theta_c\mid\alpha_c)^{\pi_c}P(\theta_d\mid\alpha)^{1-\pi_c}. \tag{3}$$

然后根据 RT 来判断文本关联关系：如果微博包含 RT，表示为一条锐推微博，转发部分 d_{RT} 与各个主题之间关系为 $\theta_{d_{\mathrm{RT}}}$，从以 λ 为参数的伯努利分布中抽取 r，来决定从参数为 $\theta_{d_{\mathrm{RT}}}$ 或 θ_d 的多项式分布中抽样出当前单词所属的主题 z_{dn}；否则 r 取 0，直接从参数为 θ_d 的多项式分布中抽样出当前单词所属的主题 z_{dn}. 最后从参数为 $\varphi_{z_{dn}}$ 的多项式分布中抽取出具体单词 w_{dn}.

一条微博中，所有单词与其所属主题的联合概率分布如式(4)所示：

$$P(w,z\mid\lambda,\theta,\beta)=P(r\mid\lambda)P(z\mid\theta)P(w\mid z,\beta)=P(r\mid\lambda)P(z\mid\theta_d)^{1-r}P(z\mid\theta_{d_{\mathrm{RT}}})^{r}P(w\mid z,\beta). \tag{4}$$

MB-LDA 中微博生成过程的形式化描述如图 3 所示：

```
for each topic k∈{1,2,…,T} do
  draw φ_k ~ Dir(β)
end for
for each microblog d do
  use "@" to choose a contact
  if starts with "@"
    draw θ_d = θ_c ~ Dir(α_c)
  else
    draw θ_d ~ Dir(α_d)
  for each word w_dn do
    use "RT" to judge a relation
    if has "RT"
      draw r = Ber(λ)
      if r=1
        draw z_dn ~ Multi(θ_{d_RT})
      else
        z_dn ~ Multi(θ_d)
    else
      draw z_dn ~ Multi(θ_d)
    draw w_dn ~ Multi(φ_{z_dn})
  end for
end for
```

Fig. 3 Generative process of microblogs.

图 3 微博的生成过程

2.3 模型的推导与主题挖掘

MB-LDA 模型的推导采用 Gibbs Sampling 的方法. Gibbs Sampling 是一种快速高效的 MCMC (Markov chain Monte Carlo)抽样方法,通过迭代抽样的方式对复杂的概率分布进行推导,多用于贝叶斯图模型的求解. MB-LDA 模型的推导过程如下:

首先用欧拉公式对式(2)展开:

$$P(w \mid z,\beta)=\left(\frac{\Gamma(V\beta)}{\Pi_v\Gamma(\beta)}\right)^T\prod_{j=1}^{T}\frac{\Pi_v\Gamma(n_{j,v}+\beta)}{\Gamma(n_{j,.}+V\beta)}; \tag{5}$$

$$P(z \mid \alpha)=\left(\frac{\Gamma(T\alpha)}{\Pi_j\Gamma(\alpha)}\right)^T\prod_{d=1}^{D}\frac{\Pi_j\Gamma(n_{d,j}+\alpha)}{\Gamma(n_{d,.}+T\alpha)}. \tag{6}$$

然后使用 Gibbs Sampling 抽样如下后验分布:

$$P(z_i=j \mid w,z_{-i},\alpha,\beta)=\frac{P(z,w \mid \alpha,\beta)}{P(z_{-i},w \mid \alpha,\beta)}$$

$$\propto\frac{n_{j,v}+\beta-1}{n_{j,.}+V\beta-1}\times\frac{n_{d,j}+\alpha-1}{n_{d,.}+T\alpha-1}. \tag{7}$$

对式(7)反复迭代,并对所有主题进行抽样,最终达到抽样结果稳定. 由于抽单词和抽主题都满足多项式分布,θ_d 和 φ_z 的结果分别如下:

$$\theta_d=\frac{n_{d,j}+\alpha-1}{n_{d,.}+T\alpha-1}; \tag{8}$$

$$\varphi_z=\frac{n_{j,v}+\beta-1}{n_{j,.}+V\beta-1}. \tag{9}$$

类似地,通过抽样得到联系人关于主题的分布 θ_c:

$$\theta_c=\frac{n_{c,j}+\alpha_c-1}{n_{c,.}+T\alpha_c-1}. \tag{10}$$

至此,MB-LDA 模型通过 Gibbs Sampling 求解出微博集中微博在主题上的概率分布 θ_d,以及主题在单词上的概率分布 φ_z. 根据 θ_d 和 φ_z 就能求出每条微博关于各个主题的概率分布以及主题关于每个单词的概率分布. 通过概率计算,对整个微博集进行分析,就可以挖掘出每条微博最可能属于某个主题、每个主题最具代表性的单词.

在微博主题挖掘之外,MB-LDA 模型还能推导出特定联系人与主题的概率分布 θ_c. 根据 θ_c 就能求出每个联系人关于各个主题的概率分布,可以挖掘出每个联系人最感兴趣的某几个主题.

综上所述,MB-LDA 模型不仅能挖掘出微博的主题,还能挖掘出联系人关注的主题. 此外,还能利用主题挖掘的结果对微博集作个性化的分析,如为特定微博寻找相似微博,为特定用户作微博推荐,帮助用户寻找感兴趣的社交圈子等等.

2.4 模型的延伸

MB-LDA 模型不仅仅可以应用于微博数据的主题挖掘,还能扩展到许多带有社交网络性质的文本中,比如 Email 数据的主题挖掘:Email 中的回复关系可以类比为微博中的转发(RT)关系,Email 的收件人、抄送人则可以类比为微博中的对话(@)关系;其他例如聊天记录的主题挖掘、论坛帖子的主题挖掘等应用场景同样也可以借鉴 MB-LDA 模型.

3 实 验

3.1 实验准备

3.1.1 数据集

本文采用的数据集[20]原始数据来源于 Twitter. 该数据集收集了 115 886 名用户于 2009 年 9 月至 2010 年 1 月发布的 3 844 612 条微博. 本文截取了其中 10 万条作为实验数据(其中包含 14 180 个联系人),使用 MB-LDA 模型对这 10 万条微博进行主题挖掘.

3.1.2 数据预处理

数据集本身包含的是原始微博数据,在使用 MB-LDA 模型分析之前必须进行一项数据预处理工作:去除停用词(stopwords). 停用词是指代词和语气助词等常用词,它们出现频率很高但对于主题挖掘没有帮助. 所以在应用 MB-LDA 模型之前要做好这项预处理工作. 本文采用停用词字典的方法将停用词去除.

3.1.3 实验环境

本文的实验环境为 Intel Pentium4 3.00 GHz 的 CPU,2 GB 的内存,160 GB 硬盘的 PC 机.操作系统为 Window XP,实验工具为 Matlab R2009b.

3.2 实验结果

3.2.1 整体效果

本文的参数设置参考文献[6]中的方法,超参数的设置为 $\alpha=\alpha_c=1$,$\beta=0.01$,$T=50$,其中 λ 默认取值为 1,表示锐推微博主题与原微博主题全相关.

MB-LDA 模型的主题挖掘整体效果如图 4 所示,共挖掘出 50 个主题.图 4 中只展示了前 6 个主题.根据各个主题对应的关键词可以发现,Topic 1 是商务相关的主题,Topic 2 是苹果产品相关的主题,Topic 3 是时间相关的主题,Topic 4 是微博相关的主题,Topic 5 是就餐相关的主题,Topic 6 是多媒体相关的主题.这说明模型挖掘到的主题,主题对应的关键词准确性较高,且主题之间的独立性较强.同时,图 4 中还展示了 Topic 2 和 Topic 6 分别对应的几个典型微博,证实微博和主题的相关性较高,划分合理(如 Topic 2 中的第 4 条微博就是通过文本关联关系找到微博和主题的联系).

在挖掘微博主题之外,图 5 还展示 MB-LDA 模型挖掘联系人和主题关系的效果.模型分析出该联系人 Ethan(化名)特别关注的几个主题为 Topic 1, Topic 4 和 Topic 6.实际与 Ethan 相关的对话微博也确实包含商务、微博或是电影视频这些主题.实验证实了 MB-LDA 模型能较好地挖掘出联系人和他们关注的主题之间的关系.

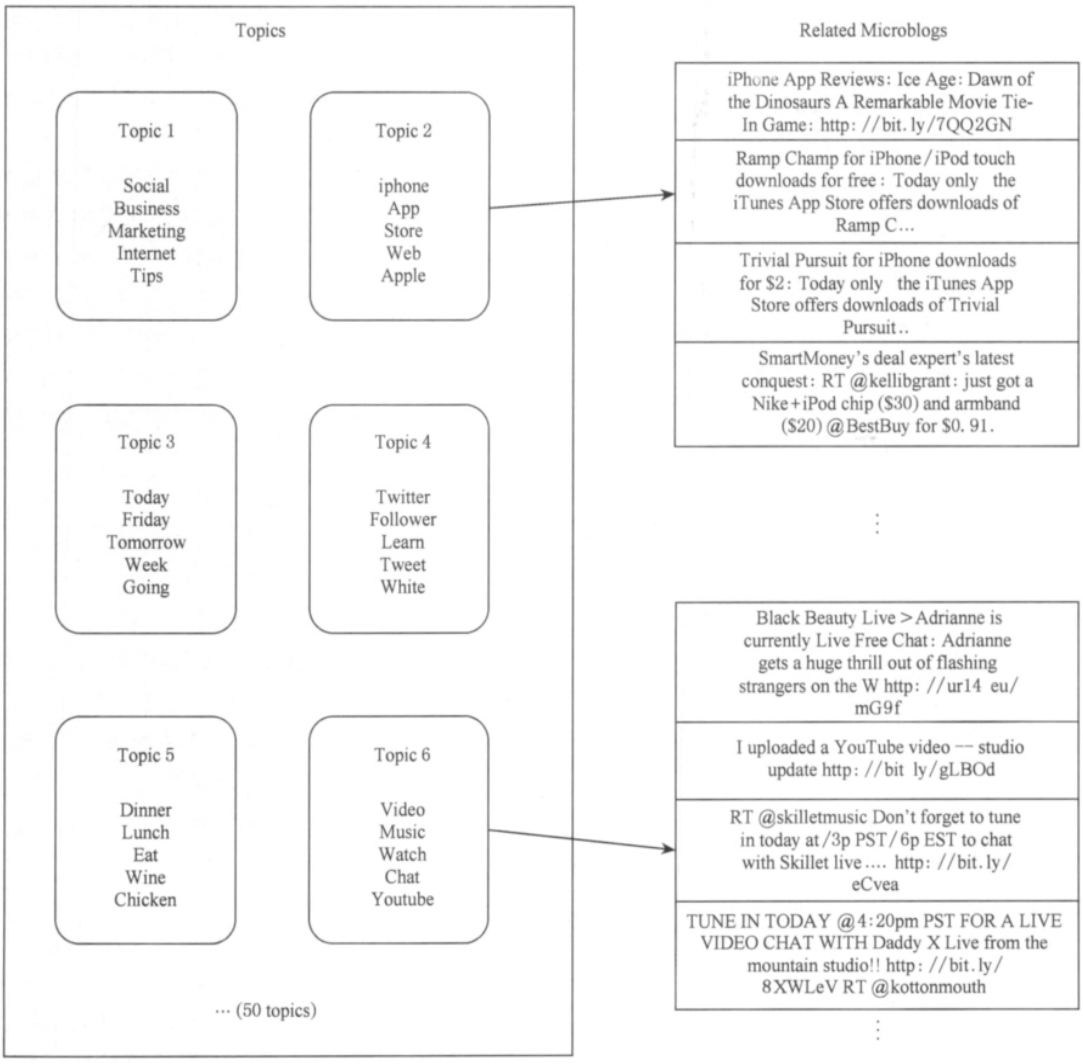

Fig. 4 Topic mining overall result of MB-LDA.

图 4 MB-LDA 模型的整体效果图

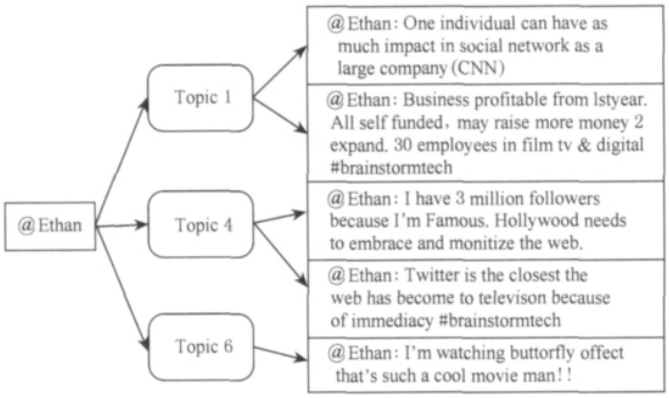

Fig. 5 Example of contactor-topic relation.

图 5 联系人对应主题的关系

3.2.2 对比实验

本文采用 *Perplexity* 指标对实验结果进行度量. *Perplexity* 是度量概率图模型性能的常用指标,也是主题建模界常用的衡量方法[9,18],表示预测数据时的不确定度,取值越小表示性能越好,模型的推广性越高. *Perplexity* 定义如下:

$$Perplexity(W)=\exp\left\{-\frac{\sum_m \ln p(w_m)}{\sum_m N_m}\right\}, \tag{11}$$

其中 W 为测试集,w_m 为测试集中可观测到的单词,N_m 为单词数.

在相同的参数设置下,通过计算 *Perplexity* 来分析模型的推广能力,计算得出 LDA 与 MB-LDA 模型的 *Perplexity* 如表 2 所示:

Table 2 *Perplexity* of LDA and MB-LDA

表 2 LDA 与 MB-LDA 模型的 *Perplexity*

Iterations	LDA	MB-LDA
50	7 092.3	6 966.6
100	6 591.3	6 495.9
150	6 361.5	6 222.7
200	6 079.2	6 007.9
250	6 049.8	6 001.3

表 2 的数据作图如图 6 所示.

同时,实验还比较了某个主题下 LDA 模型和 MB-LDA 模型的关键词差异(如图 7 所示).

通过与 LDA 模型的对比实验发现,在相同的参数条件下,随着迭代次数的增加,直到模型趋于收敛时,MB-LDA 模型的 *Perplexity* 均要小于 LDA,证明 MB-LDA 模型利用联系人关联关系和文本关联关系对微博进行分析,确实能够提高模型的性能和推广性. 同时 MB-LDA 模型得到的主题也基本与 LDA 模型相当,关键词准确性不低于 LDA 模型.

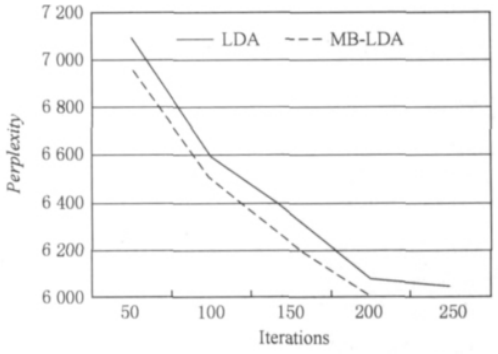

Fig. 6 Comparison of *Perplexity* of two models.

图 6 模型的 *Perplexity* 对比图

LDA:	Topic	Prob.	MB-LDA:	Topic	Prob.
	iphone	0.043 58		iphone	0.042 87
	app	0.026 80		app	0.026 36
	store	0.025 51		store	0.024 87
	download	0.023 69		web	0.016 36
	apple	0.016 48		apple	0.016 21
	support	0.015 26		windows	0.013 82
	mixtape	0.014 27		mac	0.013 59
	windows	0.014 05		version	0.012 55
	mac	0.013 82		mobile	0.012 40
	version	0.012 83		view	0.011 73

Fig. 7 Comparison of key words in the same topic.

图 7 相同主题下模型间关键词对比图

综上所述,MB-LDA 由于在模型设计中综合考虑了微博中的结构化数据(联系人信息和锐推信息)和非结构化数据(文本信息),不仅能挖掘出微博的

主题,还能挖掘出联系人关注的主题,*Perplexity* 指标优于传统的 LDA 模型,整体效果较好.

4 结束语

本文针对微博特殊的文本结构,综合考虑了微博的联系人关联关系和文本关联关系,提出了一个适合于微博主题挖掘的模型 MB-LDA. 模型不仅能准确地挖掘出微博的主题,还能有效地挖掘出联系人关注的主题.

今后的研究工作中将继续优化 MB-LDA 模型的效率,探索实时对微博数据进行主题分类,并考虑大规模数据处理等情况.

参 考 文 献

[1] Kang J H, Lerman K, Plangprasopchok A. Analyzing Microblogs with affinity propagation [C] //Proc of the 1st KDD Workshop on Social Media Analytic. New York: ACM, 2010: 67-70

[2] Ramage D, Dumais S, Liebling D. Characterizing microblogs with topic models [C] //Proc of Int AAAI Conf on Weblogs and Social Media. Menlo Park, CA: AAAI, 2010: 130-137

[3] Xu R, Wunsch D. Survey of clustering algorithms [J]. IEEE Trans on Neural Networks, 2005, 16(3): 645-678

[4] Deerwester S, Dumais S, Landauer T, et al. Indexing by latent semantic analysis [J]. Journal of the American Society of Information Science, 1990, 41(6): 391-407

[5] Landauer T K, Foltz P W, Laham D. Introduction to Latent Semantic Analysis [J]. Discourse Processes, 1998, 25(2): 259-284

[6] Griffiths T, Steyvers M. Probabilistic topic models [G] // Latent Semantic Analysis: A Road to Meaning. Hillsdale, NJ: Laurence Erlbaum, 2006

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

[8] Salton G, McGill M. Introduction to Modern Information Retrieval [M]. New York: McGraw-Hill, 1983

[9] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. The Journal of Machine Learning Research, 2003, 3: 993-1022

[10] Wei X, Croft W B. LDA-based document models for ad hoc retrieval [C] //Proc of the 29th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2006: 178-185

[11] Dietz L, Bickel S, Scheffer T. Unsupervised prediction of citation influences [C] //Proc of the 24th Int Conf on Machine Learning. New York: ACM, 2007: 233-240

[12] Mei Qiaozhu, Cai Deng, Zhang Duo, et al. Topic modeling with network regularization [C] //Proc of the 17th Int Conf on World Wide Web. New York: ACM, 2008: 101-110

[13] Blei D M, Lafferty J. Text Mining: Classification, Clustering, and Applications [M]. New York: Chapman & Hall/CRC, 2009

[14] Blei D M, Lafferty J D. Dynamic topic models [C] //Proc of the 23rd Int Conf on Machine Learning. New York: ACM, 2006: 113-120

[15] Boyd-Graber J, Blei D M, Syntactic topic models [C] //Proc of the 20th Neural Information Processing Systems (NIPS). Cambridge: MIT, 2008

[16] Nallapati R, Cohen W. Link-pLSA-LDA: A new unsupervised model for topics and influence of blogs [C] // Proc of the Int Conf on Weblogs and Social Media (ICWSM). Menlo Park, CA: AAAI, 2008

[17] Sun C, Gao Bin, Cao Zhenfu, et al. HTM: A topic model for hypertexts [C] //Proc of the 2008 Conf on Empirical Methods in Natural Language Processing. New York: ACM, 2008: 514-522

[18] Griffiths T L, Steyvers M. Finding scientific topics [J]. Proc of the National Academy of Sciences of the United States of America, 2004, 101(Suppl1): 5228-5235

[19] Minka T P, Lafferty J. Expectation-propagation for the generative aspect model [C] //Proc of the 18th Conf on Uncertainty in Artificial Intelligence. Boston, MA: AUAI, 2002: 352-359

[20] Cheng Z, Caverlee J, Lee K. You are where you tweet: A content-based approach to geo-locating twitter users [C] // Proc of the 19th ACM Int Conf on Information and Knowledge Management (CIKM). New York: ACM, 2010: 759-768

Zhang Chenyi, born in 1987. PhD candidate in the College of Computer Science and Technology, Zhejiang University. His current research interests include data mining, topic model and large scale unstructured data management.

Sun Jianling, born in 1964. Professor of the College of Computer Science and Technology, Zhejiang University. His main research interests include database, distributed system and finance information engineering.

Ding Yiqun, born in 1980. PhD, researcher of Industrial Technology Research Institute of Zhejiang University. His main research interests include Bayesian model, Monte Carlo approximate probabilistic inference and text mining.