基于MRT-LDA模型的微博文本分类

庞雄文 万本帅 王 盼
(华南师范大学计算机学院 广州 510631)

摘 要 微博的广泛使用产生了大量微博数据,这些数据中包含有大量有价值的信息。然而由于微博信息的文本内容简短且其本身带有一些结构化的社会网络方面的信息,传统的主题模型建模方法并不能十分有效地处理微博信息。根据微博信息的特点,提出一个基于Latent Dirichlet Allocation (LDA)的微博生成模型MRT-LDA,利用微博之间的转发、对话、支持(赞)和评论等关系来计算微博之间的相关性,综合考虑微博之间的相关性和同一用户微博信息间的关系,来辅助对微博的主题进行挖掘。采用吉布斯抽样法对模型进行推导,结果表明该模型能有效地对微博数据进行文本挖掘。

关键词 微博,主题挖掘,LDA,MRT-LDA,概率生成模型,社交网络
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.08.040

Micro-blog's Text Classification Based on MRT-LDA

PANG Xiong-wen WAN Ben-shuai WANG Pan
(School of Computer, South China Normal University, Guangzhou 510631, China)

Abstract Micro-blog's widespread use has produced a large number of micro-blog data, which contains a large number of valuable information. However, due to the short text content of micro-blog information and its own information on the social network, the traditional model method is not so effective to deal with micro-blog information. For this kind of special text, the traditional text mining algorithm can't be very good. Based on latent dirichlet allocation (LDA), this paper put forward a micro blogging generation model MRT-LDA according to the characteristics of micro blog information, which takes the relations between Chinese micro-blog documents and other Chinese micro-blog documents into consideration to help topic mining in micro-blog. Gibbs sampling method is used to inference the model, the results indicate that the model can offer an effective solution to text mining for Chinese micro-blog.

Keywords Micro-blog, Topic mining, LDA, MRT-LDA, Probabilistic generative model, Social network

1 引言

微博作为一种互联网社交网络服务,以快速、便捷的特性被广泛应用。基于用户之间的关联关系,微博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。用户通过微博网站构建的平台可以发布自己最新的状态或表达自己对事物的观点,也可以对某人单独发起对话,还可以转发别人的微博。

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,但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,每条微博可被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。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。

目前对微博的研究方法很少有只针对微博发布内容的,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,没有考虑微博文本内蕴含的特殊的结构化信息和微博文本简短的特性,不能很好地对微博数据进行建模。本文在研究LDA的基础上,将微博的结构化信息和微博的短文本特性考虑在内,提出了一种基于微博间关系的生成模型MRT-LDA,并利用该模型在不同的数据集上进行了仿真实验。

2 相关工作

近年来,文本主题挖掘受到了人们广泛的关注和研究,各类算法不断涌现,这些算法有很多被应用到了微博挖掘方面。Giovanni等[1]提出了一种基于相关时间序列的相似性方法对微博进行聚类。Lazard A J[2]应用文本分析的方法对具有沟

到稿日期:2016-07-04 返修日期:2016-10-19 本文受国家科技部项目(2015BAK36B06)资助。
庞雄文(1972-),男,副教授,硕士生导师,主要研究方向为机器学习和大数据技术,E-mail:augeang@163.con;万本帅(1987-),男,硕士生,主要研究方向为数据挖掘;王 盼(1992-),女,硕士生,主要研究方向为大数据技术。

通互动的 Tweets 进行信息挖掘。杨伟等人[3]应用文本挖掘和地理空间方法来检测气候和季节性对 twitter 用户的抑郁情绪的影响。Bouazizi 等人[4]提出了一种利用最小特征集对微博进行分类的方法。除此之外，主体模型 LDA 也被广泛用于大规模文档集的主题挖掘。

主题模型(Topic Model)是基于概率的生成式模型。主题模型中认为主题能够利用一定的策略方法产生单词，文档能够利用一定的策略方法产生主题。因为文档的词分布情况是已知的，所以可以利用概率的手段逆推得到文档集的主题概率分布情况。比较经典的主题模型包括 PLSA(Probabilistic Latent Semantic Analysis)和 LDA。PLSA 是 Haffman[5]在潜在语义分析(LSA)的基础上提出的概率统计分析算法。LSA 是以线性代数为基础而提出的处理文本主题的新算法，与一般的基于空间向量和统计语言模型的方法不同。传统的语言模型是直接处理文档中的单词，LSA 则在语义方面进行了一定的处理并提出了语义维度的概念。PLSA 把最大似然估计法和产生式模型考虑进 LSA 中，以提高模型的性能。PLSA 同样使用 LSA 中“降维”的思路，通过文档表示方法 TF. IDF[6]建模后，使得文档变为一种高维数据。但是主题的数目一般是比较少的，对应着低维度的语义空间。主题挖掘一般是利用降维方法将高维度的语义空间变换到一个低维度的语义空间。PLSA 一般采用 EM[7](Estimation-Maximization)算法作为模型的推理方法。因为 EM 算法的时间复杂度比 SVD 算法的时间复杂度好，所以 PLSA 在应用于海量数据领域时其性能也往往比 LSA 高。PLSA 也存在不足之处，其文档、语义和词之间的关系是由参数实现的，过多的参数会使得 PLSA 发生“过拟合现象”。

鉴于 PLSA 存在的不足，Blei 在研究 PLSA 模型的基础上提出了 LDA[8]模型。LDA 是一个概率生成算法，其过程是模拟文本的生成过程，并且是一种非监督的学习方法。LDA 和 PLSA 的关系很密切，它是在 PLSA 模型的基础上引入了 Dirichlet 先验分布的概念，而且将超参数也考虑到模型中。可以认为 LDA 是“文档-主题-词语”三层贝叶斯模型。LDA 利用已有的训练集中的观察变量，通过逆推过程得到每个文档的主题分布情况。LDA 模型被提出后，越来越多的研究者参与到主题模型的应用与改进中。以概率统计方法的知识来理解，主题是词的一种概率分布，由词汇的分布来对主题进行语义的表达；而且 Topic 是与语料库息息相关的，不一样的语料集训练得到的 Topic 也会不一样。主题模型其实就是语言模型，可对文本建模得到文本中主题的分布情况；主题模型还可以对各种各样的文本建立对应的模型。主题是词的概率分布，因此在应用中使用者一般会选择概率比较大的几个来代表主题的含义。

主题模型因为是依据数学知识建立起来的，而且比较容易进行扩展改进，所以被广泛应用于主题挖掘[8-10]、文本分类[11-12]、引文分析[13]等领域，同时也可以应用于图形处理和计算机视觉[14-15]。LDA 模型是最简单和经典的主题模型，被越来越多的研究者应用，许多对于 LDA 模型的扩展和变形也相继被提出。如，为了解决主题的孤立问题，Blei[16]提出了 HLDA (Hierarchical LDA)模型；为了更好地处理文本和主题的表达，利用文档存在的标签，Ramage D 等人[17]提出了适合处理带有标签的文档数据集的 Labeled LDA；为了让 LDA 模型能够进行动态调整，Blei[18]又将时间考虑进 LDA 模型中，从而提出 Dynamic Topic Model。

LDA 在微博信息处理方面也出现了一些应用和改进，但大都是针对 Tweets 进行的。基于 Tweets 中用“#”符号标示作为标签，Ramage D 等人[19]将 Labeled LDA 用于处理分析 Twitter 数据。但是 Labeled LDA 对中文微博的处理效果并不十分理想，因为中文微博中较少有用“#”符号标示的信息。也有学者直接把 LDA 应用于 Twitter 文本数据的处理，如 Pennacchiotti[20]把 LDA 应用于 Twitter 的分类，Weng[21]利用 LDA 发现有影响的用户。由于微博文本比传统文本要短得多，因此 LDA 在处理短文本时效果会有所降低。为了应对文本内容简短的问题，Pennacchiotti[20]，Weng[21]和 Hong[22]等提出把一个用户的所有信息作为一个文本。Zhao[23]提出了 Twitter-LDA 模型，Twitter-LDA 模型认为一个 Twitter 信息包括一个主题且 Twitter 信息包括主题词和背景词。该模型主要针对单个 Twitter 信息进行考虑，没有考虑 Twitter 用户之间的影响。Iwata[24]提出了 TTM 模型，该模型主要是考虑了 Twitter 信息的时序性，把 Twitter 信息分成各个时间段，能够更好地动态更新模型。Sasaki K[25]通过将 Twitter-LDA 模型和 TTM 模型结合形成了 Twitter-TTM 模型，该模型结合了 Twitter-LDA 模型和 TTM 模型的优点，但是也未考虑微博之间的联系性。张晨逸[26]在研究主题模型 LDA 的基础上提出了一个新的微博文本模型 MB-LDA，其综合利用了微博信息的转发关系和对话(@)关系，来辅助对微博信息的主题挖掘。但 MB-LDA 模型没有解决文本简短的问题，并且只考虑了转发关系和@关系，没有涉及不存在转发关系和@关系的情况。另外，MB-LDA 主要是针对 Twitter 信息的。李敬[27]利用微博信息中的@、转发和话题标签，把微博划分成用户兴趣、用户互动和话题微博 3 种类别，在 ATM 的基础上提出了一个改进的主题模型 HC-ATM。陶永才[28]提出了基于转发的狄利克雷分配(RT-LDA)模型，该模型考虑了微博转发关系。本文针对微博之间的转发、对话(@)、点赞和评论等关系，考虑了微博之间的相互影响，提出了一个统一建模微博各种关系的 MRT-LDA 模型，该模型能够更好地利用微博的不同关系来提高主题挖掘的性能。

3　基于微博间关系的 MRT-LDA 模型

本文按时间片把每个用户的多个微博信息合成一个文本，并根据同一个用户相邻时间片间的微博信息的联系和微博用户之间的转发、对话(@)、点赞和评论等关系，提出了一个统一建模微博各种关系的 MRT-LDA 模型，其能够更好地利用微博间的不同关系来提高数据挖掘的性能。

3.1　MRT-LDA 模型的文本内容扩展

不同于一般的文本，微博文本是社交类文本，具有文本简短以及联系性的特点。

为了解决文本简短的问题，采用按时间片划分的方式扩

充文本，即根据时间片把多条微博合为一个微博文本。具体思路如下：

(1)设置时间片大小；

(2)按时间片把多条微博合为一个文本。

3.2 MRT-LDA 模型的文本关系改进

为了更具体地分析文本之间的联系，考虑微博文本之间的@关系、转发关系、评论和点赞等，并综合这些关系提出了一个表达文本之间的联系性因子。表达式如下：

$$MBR=w_1\cdot r_1+w_2\cdot r_2+w_3\cdot r_3+w_4\cdot r_4 \tag{1}$$

其中，r_1 表示在文本之间有无转发的关系，r_2 表示有无@关系，r_3 表示有无评论关系，r_4 表示有无点赞关系，w_1，w_2，w_3，w_4 分别为 r_1，r_2，r_3，r_4 对应的权重。

本文利用关联性因子公式求出每对文本之间的关系性因子值 MBR。当微博文本之间的关系性因子值 MBR 为 0 时，认为文本在生成时相互之间没有影响；MBR 不为 0 时，认为微博之间可能会存在关联性关系。若 MBR 不为 0，则根据 MBR 的值设定一个 s 值。本文假设两个微博文本是否存在关系服从二项分布，s 为此二项分布的参数。从该二项分布中抽取一个值 r，若 r 为 0，则认为不清楚两微博文本之间是否存在关系；若 r 为 1，则认为两个微博文本之间存在关系。

MRT-LDA 是在研究 LDA 的基础上，把文档关系作为一个影响因子进行建模而提出的适用于处理微博文本的模型。

MRT-LDA 的贝叶斯网络图如图 1 所示。其中，r_i 是文本关系，服从二项分布，其参数为 s_i，文本关系性因子 MBR 决定 s_i 的值，表示文本 d 和文本 d_i 中的文本关系信息。根据各个文本间 r_i 的值来确定与生成文本相关的文本。若 $r_i=1$，则表示文本之间有关系，否则就不考虑文本间的关系。首先从参数为 β 的 Dirichlet 分布中为每个主题抽取主题与单词的概率分布 φ。在生成文本时，若该文本没有相关文本，则从参数为 α_t 的 Dirichlet 分布中为该文本抽取一个文本与主题的概率分布 ϑ。α_t 是 $t-1$ 时的文本主题分布 ϑ_d^{t-1}。若存在相关文本 $d_1,d_2,\cdots,d_n$，则从参数为 α_d 的 Dirichlet 分布中为该文本抽取文本与主题的概率分布 ϑ_d，α_d 由相关文本 $d_1,d_2,\cdots,d_n$ 的主题分布以及该用户 $t-1$ 时间片的文本主题分布决定。之后根据文本的主题概率分布 θ_d 为文本中的每个词选择主题，再根据对应的主题与词的分布抽取词，对每个词重复上述过程即得到了一篇文本。

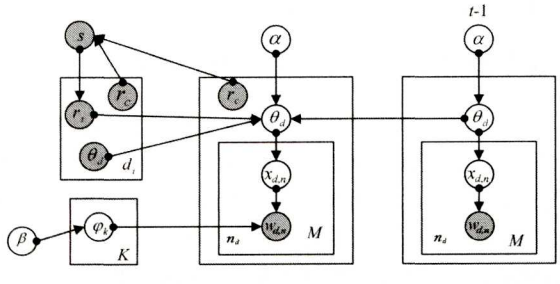

图 1 MRT-LDA 的贝叶斯网络图

在数据集中，一个微博文本的文本主题分布 ϑ_d 的概率分布为：

$$p(\vartheta_d|\alpha,\mathrm{e},s)=\{(1-\mathrm{e})+\mathrm{e}(1-r)p(r|s)\}p(\vartheta_d|\alpha_t)+\mathrm{e}\times r\times(r|s)p(\vartheta_d|\alpha_d) \tag{2}$$

$$\alpha_d=\frac{1}{1+W_t}\alpha_h\left(W_t\vartheta_d^{t-1}+\frac{\sum_{i=0}^{d_n}r_{di}\vartheta_{di}}{\sum_{i=0}^{d_n}r_{di}}\right) \tag{3}$$

其中，$\vartheta_{d1},\vartheta_{d2},\cdots,\vartheta_{di},\cdots,\vartheta_{dn}$ 为文档 d 的相关文本的主题分布，α_h 表示矩阵 α 各元素之和，α_d 是由 ϑ_{di} 以及 ϑ_d^{t-1} 的加权均值再乘以 α 得到的，r_{di} 表示当前微博文本与文本 d_i 是否有关系，$p(r|s)$ 表示文本存在关系文本的概率。

在一篇微博文本中，所有的词和其所属主题的联合概率分布如下式所示：

$$\begin{aligned}&p(W,Z|\alpha_t,\beta,\mathrm{e},s)\\&=p(W|Z,\beta)p(Z|\vartheta_d)p(\vartheta_d|\alpha_t,\mathrm{e},s)\\&=p(W|Z,\beta)p(Z|\vartheta_d)\times\{\{(1-\mathrm{e})+\mathrm{e}\times p(r|s)(1-r)\}\\&\quad p(\vartheta_d|\alpha_t)+\mathrm{e}\times p(r|s)\times r\times p(\vartheta_d|\alpha_d)\}\end{aligned} \tag{4}$$

MRT-LDA 模型的生成过程如下。

对于每篇微博文档 d：

(1)根据参数 β 的 Dirichlet 分布为每个主题选择词分布，抽取 $\varphi\sim\mathrm{Dir}(\beta)$。

(2)计算该文档与其他文档的关系值，确定参数 s_i，抽取 $r_i\sim\mathrm{B}(1,s_i)$。

(3)若存在 $r_i=1$，则根据相关文本的主题分布 $\vartheta_{d1},\vartheta_{d2},\cdots,\vartheta_{di},\cdots,\vartheta_{dn}$ 和同一用户上一时间片文本的主题分布 ϑ_d^{t-1} 计算出 α_d，选择该文档主题分布 $\vartheta_d\sim\mathrm{Dir}(\alpha_d)$；否则，从参数为 α_t 的 Dirichlet 分布中抽取一个主题分布 $\vartheta_d\sim\mathrm{Dir}(\alpha_t)$。

(4)对于该文档的每个词 $W_{d,n}$：

1)抽取一个主题 $Z_{d,n}$。

2)根据主题 $Z_{d,n}$ 对应的词分布抽取一个词 $W_{d,n}$。

4 模型实现

4.1 MRT-LDA 模型的推导

MRT-LDA 模型采用近似推理中的吉布斯抽样方法进行推导。吉布斯抽样是近似推理，用其对 MRT-LDA 模型中的隐含变量学习估计，不但更加简单易懂，而且可以得到较为理想的效果。该算法的执行方式是每次选择概率向量的一个维度，并根据已知的其他维度的变量值选取当前维度的值，一直迭代到参数收敛，最后输出待估计的参数。

Gibbs Sampling 中的抽样迭代公式为：

$$\begin{aligned}p(z_i=k|z_{\neg i},W)&=\frac{p(z,W)}{p(z_{\neg i},W)}\\&=\frac{p(W|z)}{p(W_{\neg i}|z_{\neg i})p(W_i)}\frac{p(z)}{p(z_{\neg i})}\end{aligned} \tag{5}$$

又因为 $p(W|z)$ 可以化为：

$$\begin{aligned}p(W|z)&=\int p(W|z,\varphi)\ p(\varphi|\beta)\mathrm{d}\varphi\\&=\int\prod_{z=1}^{K}\prod_{t=1}^{V}\varphi_{z,t}^{n_z^t}\cdot\prod_{z=1}^{K}\frac{1}{\Delta(\beta)}\prod_{t=1}^{V}\varphi_{z,t}^{\beta_t-1}\mathrm{d}\varphi_z\\&=\int\prod_{z=1}^{K}\frac{1}{\Delta(\beta)}\prod_{t=1}^{V}\varphi_{z,t}^{n_z^t+\beta_t-1}\mathrm{d}\varphi_z\end{aligned} \tag{6}$$

且 $\int\frac{1}{\Delta(\beta)}\prod_{t=1}^{V}\varphi_{z,t}^{n_z^t+\beta_t-1}\mathrm{d}\varphi_z=\Delta(n_z+\beta)$，因此得到：

$$p(W|z)=\prod_{z=1}^{K}\frac{\Delta(n_z+\beta)}{\Delta(\beta)} \tag{7}$$

同理可得：

$$p(W_{\neg i}|z_{\neg i})=\prod_{z=1}^{K}\frac{\Delta(n_{z,\neg i}+\beta)}{\Delta(\beta)} \tag{8}$$

$$p(z)=\prod_{m=1}^{M}\frac{\Delta(n_m+\alpha)}{\Delta(\alpha)} \tag{9}$$

$$p(z_{\neg i})=\prod_{m=1}^{M}\frac{\Delta(n_{m,i}+\alpha)}{\Delta(\alpha)} \tag{10}$$

又因为除第 i 个词所属的文档和主题外，其他文档的主题分布都没变，且 $p(W_i)$ 是常量，$\Gamma(n)=(n-1)!$，所以：

$$\begin{aligned}
&p(z_i=k|z_{\neg i},W)\\
&=\frac{\prod_{z=1}^{K}\frac{\Delta(n_z+\beta)}{\Delta(\beta)}}{\prod_{z=1}^{K}\frac{\Delta(n_{z,\neg i}+\beta)}{\Delta(\beta)}}p(W_i)\frac{\prod_{m=1}^{M}\frac{\Delta(n_m+\alpha)}{\Delta(\alpha)}}{\prod_{m=1}^{M}\frac{\Delta(n_{m,\neg i}+\alpha)}{\Delta(\alpha)}}\\
&\propto\frac{\frac{\Delta(n_k+\beta)}{\Delta(\beta)}}{\frac{\Delta(n_{k,\neg i}+\beta)}{\Delta(\beta)}}\frac{\frac{\Delta(n_{di}+\alpha)}{\Delta(\alpha)}}{\frac{\Delta(n_{di,\neg i}+\alpha)}{\Delta(\alpha)}}\\
&=\frac{\Delta(n_k+\beta)}{\Delta(n_{k,\neg i}+\beta)}\frac{\Delta(n_{di}+\alpha)}{\Delta(n_{di,\neg i}+\alpha)}\\
&=\frac{\prod_{t=1}^{V}\Gamma(n_k^t+\beta_t)}{\Gamma(\sum_{t=1}^{V}(n_k^t+\beta_t))}\frac{\prod_{k=1}^{K}\Gamma(n_{di}^k+\alpha_{t,k}^{1-r}\cdot\alpha_{d,k}^r)}{\Gamma(\sum_{k=1}^{K}(n_{di}^k+\alpha_{t,k}^{1-r}\cdot\alpha_{d,k}^r))}\\
&\quad\frac{\Gamma(\sum_{t=1}^{V}(n_{k,\neg i}^t+\beta_t))}{\prod_{t=1}^{V}\Gamma(n_{k,\neg i}^t+\beta_t)}\frac{\Gamma(\sum_{k=1}^{K}(n_{di,\neg i}^k+\alpha_{t,k}^{1-r}\cdot\alpha_{d,k}^r))}{\prod_{k=1}^{K}\Gamma(n_{di,\neg i}^k+\alpha_{t,k}^{1-r}\cdot\alpha_{d,k}^r)}\\
&=\frac{(n_k^t+\beta_t)-1}{\sum_{t=1}^{V}(n_k^t+\beta_t)-1}\frac{(n_{di}^k+\alpha_{t,k}^{1-r}\cdot\alpha_{d,k}^r)-1}{\sum_{k=1}^{K}(n_{di}^k+\alpha_{t,k}^{1-r}\cdot\alpha_{d,k}^r)-1}\\
&=\frac{(n_k^t+\beta_t)-1}{\sum_{t=1}^{V}(n_k^t+\beta_t)-1}\frac{(n_{di}^k+\alpha_{t,k}^{1-r}\cdot\alpha_{d,k}^r)-1}{\sum_{k=1}^{K}(n_{di}^k+\alpha_{t,k}^{1-r}\cdot\alpha_{d,k}^r)-1}
\end{aligned} \tag{11}$$

其中，z_i 表示第 i 个单词分配的主题，$z_{\neg i}$ 表示除第 i 个单词以外所有词的主题分布情况，v 表示字典中词的个数，$n_{k,\neg i}^t$ 表示除第 i 个词以外单词 t 分配给主题 k 的次数，r 表示是否有相关的微博文本。

下面通过计算来研究 MRT-LDA 模型中 φ 和 ϑ 的性质。

$$\begin{aligned}
p(\varphi_k|z,W,\beta)&=\frac{p(W|\varphi_k)p(\varphi_k|z,\beta)}{\int p(W|\varphi_k)p(\varphi_k|z,\beta)\mathrm{d}\varphi_k}\\
&=\frac{\prod_{t=1}^{n_v}\varphi_{k,t}^{n_k^t}\frac{1}{\Delta(\beta)}\prod_{t=1}^{n_v}\varphi_{k,t}^{\beta_t-1}}{\int\prod_{t=1}^{n_v}\varphi_{k,t}^{n_k^t}\frac{1}{\Delta(\beta)}\prod_{t=1}^{n_v}\varphi_{k,t}^{\beta_t-1}d\varphi_k}\\
&=\frac{\prod_{t=1}^{n_v}\varphi_{k,t}^{n_k^t}\prod_{t=1}^{n_v}\varphi_{k,t}^{\beta_t-1}}{\int\prod_{t=1}^{n_v}\varphi_{k,t}^{n_k^t}\prod_{t=1}^{n_v}\varphi_{k,t}^{\beta_t-1}\mathrm{d}\varphi_k}\\
&=\frac{\prod_{t=1}^{n_v}\varphi_{k,t}^{n_k^t+\beta_t-1}}{\int\prod_{t=1}^{n_v}\varphi_{k,t}^{n_k^t+\beta_t-1}\mathrm{d}\varphi_k}
\end{aligned} \tag{12}$$

又因为 $\int\prod_{t=1}^{n_v}\varphi_{k,t}^{n_k^t+\beta_t-1}\mathrm{d}\varphi_k=\Delta(n_k+\beta)$，所以式(12)可以变成 $p(\varphi_k|z,W,\beta)=\mathrm{Dir}(\varphi_k|n_k+\beta)$。

$P(\vartheta_{di}|z,\alpha)$ 的计算公式如下：

$$\begin{aligned}
&P(\vartheta_{di}|z,\alpha)\\
&=\frac{P(z|\vartheta_{di})P(\vartheta_{di}|\alpha)}{\int P(z|\vartheta_{di})P(\vartheta_{di}|\alpha)\mathrm{d}\vartheta_{di}}\\
&=\frac{\prod_{n=1}^{N_{di}}P(z_n|\vartheta_{di})P(\vartheta_{di}|\alpha_t^{1-r}\cdot\alpha_d^r)}{\int\prod_{n=1}^{N}P(z_n|\vartheta_{di})P(\vartheta_{di}|\alpha_t^{1-r}\cdot\alpha_d^r)\mathrm{d}\vartheta_{di}}\\
&=\frac{\prod_{k=1}^{K}\vartheta_{di,k}^{n_{di}^k}\frac{1}{\Delta(\alpha^{1-r}\cdot\alpha_d^r)}\prod_{k=1}^{K}\vartheta_{di,k}^{\alpha_{t,k}^{1-r}\cdot\alpha_{d,k}^r-1}}{\int\prod_{k=1}^{K}\vartheta_{di,k}^{n_{di}^k}\frac{1}{\Delta(\alpha^{1-r}\cdot\alpha_d^r)}\prod_{k=1}^{K}\vartheta_{di,k}^{\alpha_{t,k}^{1-r}\cdot\alpha_{d,k}^r-1}\mathrm{d}\vartheta_{di}}\\
&=\frac{\prod_{k=1}^{K}\vartheta_{di,k}^{n_{di}^k}\prod_{k=1}^{K}\vartheta_{di,k}^{\alpha_{t,k}^{1-r}\cdot\alpha_{d,k}^r-1}}{\int\prod_{k=1}^{K}\vartheta_{di,k}^{n_{di}^k}\prod_{k=1}^{K}\vartheta_{di,k}^{\alpha_{t,k}^{1-r}\cdot\alpha_{d,k}^r-1}\mathrm{d}\vartheta_{di}}\\
&=\frac{\prod_{k=1}^{K}\vartheta_{di,k}^{n_{di}^k+\alpha_{t,k}^{1-r}\cdot\alpha_{d,k}^r-1}}{\int\prod_{k=1}^{K}\vartheta_{di,k}^{n_{di}^k+\alpha_{t,k}^{1-r}\cdot\alpha_{d,k}^r-1}\mathrm{d}\vartheta_{di}}
\end{aligned} \tag{13}$$

又因为 $\int\prod_{k=1}^{K}\vartheta_{di,k}^{n_{di}^k+\alpha_{t,k}^{1-r}\cdot\alpha_{d,k}^r-1}\mathrm{d}\vartheta_{di}=\Delta(n_{di}+\alpha_t^{1-r}\cdot\alpha_d^r)$，所以式(13)可以变为 $P(\vartheta_{di}|z,\alpha)=\mathrm{Dir}(\vartheta_{di}|n_{di}+\alpha_t^{1-r}\cdot\alpha_d^r)$。

当吉布斯抽样收敛后，需要根据文档集中所有单词的主题分配来计算 φ 和 ϑ，其计算公式如下：

$$\vartheta_{m,k}=\frac{n_{m,k}+\alpha_{t,k}^{1-r}\alpha_{d,k}^r}{\sum_{k=1}^{K}(n_{m,k}+\alpha_{t,k}^{1-r}\alpha_{d,k}^r)} \tag{14}$$

$$\varphi_{k,t}=\frac{n_{k,t}+\beta_t}{\sum_{t=1}^{v}(n_{k,t}+\beta)} \tag{15}$$

4.2 文本分类

MRT-LDA 模型可以通过吉布斯抽样的方法求出每个微博文本中的主题分布 ϑ_m，并且求出每个主题中词的分布情况 φ_k，结合微博文本的主题分布 ϑ_m 和主题的词概率分布 φ_k，就可以进一步挖掘该微博文本所关注的主题，也可以利用分类算法把微博文本中的主题分布 ϑ_m 作为分类器的输入用于分类。

5 模型实验

本文在真实微博数据集上验证 MRT-LDA 模型的性能，并将 MB-LDA 和 LDA 作为对比算法。本文还尝试把财经新闻之间的超链接作为关系，再将 MRT-LDA 模型用于财经新闻分类，并在真实的财经新闻数据上进行了验证。实验采用 micro-F1度量值来衡量文中所提模型的分类性能。用困惑度衡量 MRT-LDA 模型的泛化性能。在分类方面，都是通过与 SVM 结合来达到分类的效果。超参数 α 和 β 的经验取值为 $\alpha=50/K$，$\beta=0.01$。把 K 的取值设置为变量，通过改变 K 的大小调节模型的效果。

5.1 实验准备

5.1.1 微博数据集

本实验的数据是从实际微博平台中搜集到的，拟通过实

际的微博数据集来评估 MRT-LDA 模型的有效性。原始的新浪微博数据集是在新浪微博抓取的 1036 个用户从 2014 年 5 月 3 日到 6 月 9 日的 413572 条微博数据。原始的腾讯微博数据集是在腾讯微博抓取的 1108 个用户从 2014 年 5 月 3 日到 6 月 9 日的 402759 条微博数据。

在用模型进行建模之前，要对微博数据进行预处理。首先对原始微博数据作简单处理，把微博字数少于 10 且不包含转发、@、评论、赞等关系的微博删除，同时还把数据中微博条数少于 20 的用户的微博数据删除；然后使用中科院的分词工具进行中文分词，并且去掉停用词、地名以及一些类似于“有”、“在”、“我”这种不能表达主题的单个汉字，并按一天作为时间片对每个用户的微博信息进行划分。

5.1.2 财经新闻数据集

用从新浪财经、腾讯财经和搜狐财经抓取的实际数据来进行实验。在实验之前，先对数据做预处理。根据新浪财经中的分类，把抓取的新闻信息分为宏观、股票、基金、期货、黄金、债券和新三板这 7 类信息。根据腾讯财经中的分类，把抓取的新闻信息分为宏观、港股、美股、基金、理财、保险和新三板这 7 类信息。根根新浪财经和腾讯财经的实际分类，对抓取的信息进行标记分类。实验之前，还需要对文本信息进行分词和去除停用词等操作，以去掉没有实际意义的词。

5.2 实验结果

5.2.1 腾讯微博数据集

图 2 示出了当 $w_t=2$ 时，腾讯微博数据集上不同主题数目下模型的 micro-F1 的变化情况。从图 2 可以看出，MB-LDA 模型、LDA 模型和 MRT-LDA 模型随着主题数 K 的增加都会达到一个比较高的值，之后就会呈现折线形式的下降，该结果表明主题太少或太多时模型分类的效果都比较差。当主题数 K 为 100 时，LDA 模型、MB-LDA 模型的 micro-F1 值分别为 73.17%和 76.29%，而 MRT-LDA 模型达到最大值 86.35%。综合比较 LDA 模型、MB-LDA 模型和 MRT-LDA 模型的分类性能，发现 MRT-LDA 模型比 LDA 模型和 MB-LDA 模型的 micro-F1 值分别提高了 13%和 10%。

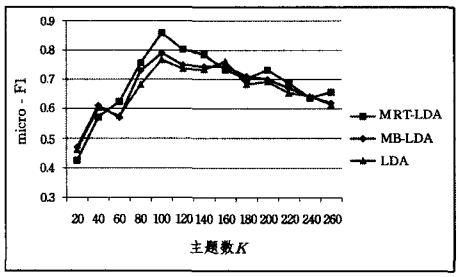

图 2 腾讯微博数据集上不同主题数目下模型的 micro-F1 值

图 3 示出了在 $K=100$ 和 $w_t=2$ 时，MB-LDA 模型、MRT-LDA 模型和 LDA 模型随着实验迭代次数的增加其 perplexity 值的变化情况。如图 3 所示，MB-LDA 模型、MRT-LDA 模型和 LDA 模型的 perplexity 值是随着迭代次数的增加而逐渐减小的，并且当模型达到一定的迭代次数之后，困惑度 perplexity 值趋于稳定。从图 3 可以看出，MRT-LDA 模型的困惑度比 MB-LDA 模型和 LDA 模型的小，说明 MRT-LDA 模型的泛化能力要优于 MB-LDA 模型和 LDA 模型。

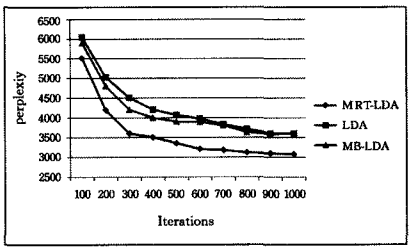

图 3 腾讯微博数据集上不同迭次数下模型的 perplexity 值

5.2.2 新浪微博数据集

图 4 示出了在新浪微博数据集上的不同主题数目下的 micro-F1 值。从图 4 可以看出，MB-LDA 模型、LDA 模型和 MRT-LDA 模型随着主题数 K 的增加其 micro-F1 值都会达到一个比较高的值，之后就会呈现折线形式的下降，该结果表示主题太少或太多时模型分类的效果都比较差。在主题数 K 为 80 时，LDA 模型和 MB-LDA 模型的 micro-F1 达到最大值 73.31%和77.51%，而 MRT-LDA 模型的 micro-F1 的最大值为 83.56%。综合比较 MB-LDA 模型、LDA 模型和 MRT-LDA 模型的分类性能，发现 MRT-LDA 模型比 LDA 模型和 MB-LDA 模型的 micro-F1 值分别提高了 10%和 6%。该结果比在腾讯微博上的结果差，可能是因为腾讯微博上的用户是基于 QQ 的，用户之间的联系会更加紧密，所以更有利于 MRT-LDA 模型性能的提高。

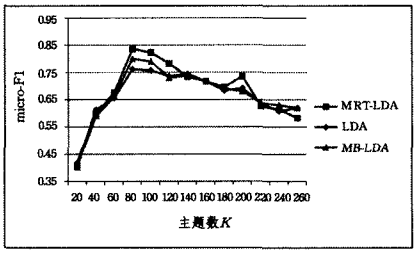

图 4 新浪微博数据集上不同主题数目下模型的 micro-F1 值

图 5 显示的是在新浪数据集上设置 $K=80$ 时，MB-LDA 模型、MRT-LDA 模型和 LDA 模型随着实验迭代次数的增加其 perplexity 值的变化情况。如图 5 所示，MB-LDA 模型、MRT-LDA 模型和 LDA 模型的 perplexity 值也是随着迭代次数的增加而逐渐减小的，且随着模型达到一定的迭代次数后，困惑度 perplexity 值趋于稳定。从图 5 中可以看出，MRT-LDA 模型的困惑度比 MB-LDA 模型和 LDA 模型的小，说明 MRT-LDA 模型的泛化能力要优于 MB-LDA 模型和 LDA 模型。在新浪数据中的困惑度 perplexity 值比腾讯微博中的低，这可能是由于新浪微博的内容更加丰富。

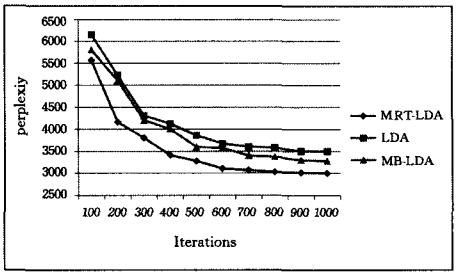

图 5 新浪微博数据集上不同迭代次数下模型的 perplexity 值

5.2.3　新浪财经数据集

图 6 示出了各模型在不同主题数目下的 micro-F1 值。从图 6 可以看出，LDA 模型和 MRT-LDA 模型随着主题数 K 的增加都是先逐渐增加到一个较高的值，之后就会逐渐下降。在主题数目比较少的情况下，模型对文本的表达效果不理想，从而导致分类性能比较低；在主题数目太多的情况下，主题之间的区分度降低，也会导致分类效果变差。SVM 的 micro-F1 值为 70.81%。MRT-LDA 模型、LDA 模型在主题数 K 为 140 时，micro-F1 值分别达到最大值 77.75%和 74.39%。MRT-LDA 模型比传统的 SVM 高 7%，比 LDA 模型提高了 3%。从整体效果来看，MRT-LDA 模型的分类性能要优于 LDA 模型，而 LDA 模型要优于传统 SVM。可见，模型在财经新闻分类中的效果比微博中的分类效果差，其原因可能是财经新闻分类是更细的分类。财经新闻的联系性比较差，导致 MRT-LDA 模型在该分类方面与 LDA 模型相比提高不多。

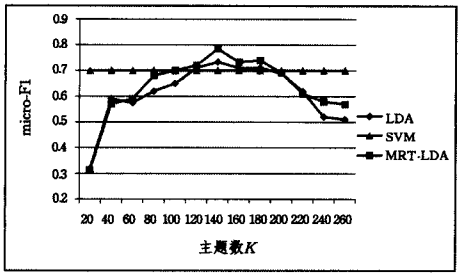

图 6　各模型在不同主题数目下的 micro-F1 值

5.2.4　腾讯财经数据集

图 7 示出了腾讯财经数据集下 SVM，LDA 和 MRT-LDA 模型的分类性能随主题数目的变化情况。LDA 模型和 MRT-LDA 模型随着主题数 K 的增加都是先逐渐增加到一个较高的值，之后就会逐渐下降。在主题数目较少的情况下，分类性能增加得较快；在 100～180 时性能达到最优。在主题数目较少的情况下，MRT-LDA 模型对文本的表达效果不理想，从而导致分类性能比较低。在主题数目太多的情况下，主题之间的区分度降低，也会导致分类效果下降。

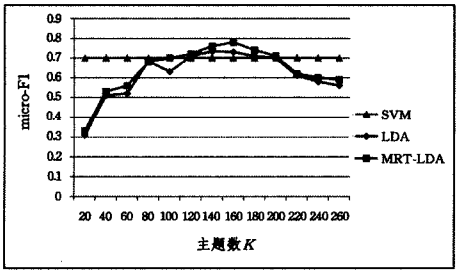

图 7　各模型在不同主题数目下的 micro-F1 值

从图 7 中的不同主题数下的整体效果来看，在 K 为 140 时，MRT-LDA 模型和 LDA 模型的分类性能到达最大值 73.56%和71.05%，SVM 的 micro-F1 值为 68.33%。MRT-LDA 模型比传统的 SVM 高 6%，比 LDA 模型提高了 3%，而 LDA 模型比传统的 SVM 高 3%。因此，MRT-LDA 模型比 LDA 模型和传统的 SVM 模型的效果都好。

在腾讯财经新闻分类中的模型的效果比新浪财经中的分类效果好，其原因可能是新浪财经新闻中的信息更丰富，也可能是因为两个数据集原本的分类标签不同。

结束语　本文在研究 LDA 模型和微博数据特点的基础上提出了 MRT-LDA 模型。首先，在新浪和腾讯微博数据集上验证了 MRT-LDA 模型的性能要优于 MB-LDA 模型和 LDA 模型 6%～13%。接着，将 MRT-LDA 模型用于财经新闻的分类中，通过实验验证了 MRT-LDA 模型比 LDA 和 SVM 的性能高出 3%～7%。综上所述，本文所提出的 MRT-LDA 模型有更好的分类性能。

MRT-LDA 模型虽然能够在一定程度上提高微博数据及关系性文本的分类效果，但是在微博信息挖掘方面还存在一些不足，微博信息中还有一些非文本信息如图片等没有被充分利用。因此在下一步工作中，考虑把这些非文本信息的因素加入到模型中，以进一步提高模型的性能。

参 考 文 献

[1] STILO G, VELARDI P. Efficient temporal mining of micro-blog texts and its application to event discovery[J]. Data Mining & Knowledge Discovery, 2016, 30(2): 372-402.

[2] LAZARD A J, SCHEINFELD E, BERNHARDT J M, et al. Detecting themes of public concern: A text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat[J]. American Journal of Infection Control, 2015, 43(10): 1109-1111.

[3] WEI Y G, LAN M, YE S. Effect of climate and seasonality on depressed mood among twitter users[J]. Applied Geography, 2015, 63: 184-191.

[4] BOUAZIZI M, OHTSUKI T. Opinion mining in Twitter: How to make use of sarcasm to enhance sentiment analysis[C]// IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. New York: ACM, 2015: 1594-1597.

[5] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis[J]. Machine Learning, 2001, 42(1): 177-196.

[6] ROELLEKE T, WANG J. TF-IDF uncovered: a study of theories and probabilities[C]// Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2008: 435-442.

[7] GEBOTYS C H, WHITE B A. EM analysis of a wireless Java-based PDA[J]. Acm Transactions on Embedded Computing Systems, 2008, 7(4): 2087-2093.

[8] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [M]. JMLR. org, 2003: 993-1022.

[9] ZHANG C Y, SUN J L. Large scale microblog mining using distributed MB-LDA[C]// Proceedings of the 21st International Conference Companion on World Wide Web. New York: ACM, 2012: 1035-1042.

[10] WEI X, CROFT W B. LDA-based document models for ad-hoc retrieval[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2006: 178-185.

（下转第 259 页）

Model Fragmentation for Distributed Execution [J]. IEEE Transactions on Systems, Man and Cybernetics-Part A: Systems and Humans, 2011, 41(2): 294-310.

[24] MEZMAZ M, MELAB N, KESSACI Y, et al. A parallel bi-objective hybrid metaheuristic for energy-aware scheduling for cloud computing systems[J]. Journal of Parallel & Distributed Computing, 2011, 71(11): 1497-1508.

[25] PRZYSTAŁKA P, KATUNIN A. Multi-Objective Meta-Evolution Method for Large-Scale Optimization Problems[M]// Recent Advances in Computational Optimization. Springer International Publishing, 2016: 165-182.

[26] KENNEDY J, EBERHART R. Particle swarm optimization [C]// IEEE International Conference on Neural Networks. IEEE, 1995: 1942-1948.

[27] JUVE G, CHERENAK A, DEELMAN E, et al. Characterizing and profiling scientific workflows[J]. Future Generation Computer System, 2013, 29(3): 682-692.

[28] DEB K, PRATAP A, AGARWAL S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II[J]. IEEE Transactions on Evolutionary Computation, 2002, 6(2): 182-197.

（上接第 241 页）

[11] MOEINZADEH H, MOHAMMADI M M, AKBARI A, et al. Evolutionary-class independent LDA as a pre-process for improving classification[C]// Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation. New York: ACM, 2009: 1909-1910.

[12] ZHANG Z F, MIAO D Q, GAO C. Short text classification using latent Dirichlet allocation [J]. Journal of Computer Applications, 2013, 33(6): 1587-1590. (in Chinese)
张志飞，苗夺谦，高灿. 基于 LDA 主题模型的短文本分类方法[J]. 计算机应用，2013，33(6)：1587-1590.

[13] DIETZ L, BICKEL S, SCHEFFER T. Unsupervised prediction of citation influences[C]// Proceedings of the 24th International Conference on Machine Learning. New York: ACM, 2007: 233-240.

[14] BLEI D M, LAFFERTY J. Text Mining: Classification, Clustering, and Applications [M]. New York: Chapman & Hall/CRC, 2009.

[15] TRUONG H P, LE T H. Fusion of bidirectional image matrices and 2D-LDA: an efficient approach for face recognition[C]// Proceedings of the Third Symposium on Information and Communication Technology. New York: ACM, 2012: 142-148.

[16] BLEI D M, JORDAN M I, GRIFFITHS T L, et al. Hierarchical Topic Models and the Nested Chinese Restaurant Process[C]// International Conference on Neural Information Processing Systems. MIT Press, 2003: 17-24.

[17] RAMAGE D, HALL D, NALLAPATI R, et al. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora[C]// Conference on Empirical Methods in Natural Language Processing: Volume. Association for Computational Linguistics, 2009: 248-256.

[18] BLEI D M, LAFFERTY J D. Dynamic topic models[C]// Proceedings of the 23rd International Conference on Machine Learning. New York: ACM, 2006: 113-120.

[19] RAMAGE D, DUMAIS S T, LIEBLING D J. Characterizing Microblogs with Topic Models[C]// International Conference on Weblogs and Social Media, LSWSM 2010. DBLP, 2010: 130-137.

[20] PENNACCHIOTTI M, POPESCU A M. A Machine Learning Approach to Twitter User Classification[J]. ICWSM, 2011, 11(1): 281-288.

[21] WENG J, LIM E P, JIANG J, et al. Twitterrank: finding topic-sensitive influential twitterers[C]// Proceedings of the Third ACM International Conference on Web Search and Data Mining. New York: ACM, 2010: 261-270.

[22] HONG L, DAVISON B D. Empirical study of topic modeling in twitter[C]// Proceedings of the first workshop on social media analytics. New York: ACM, 2010: 80-88.

[23] ZHAO W X, JIANG J, WENG J, et al. Comparing twitter and traditional media using topic models[M]// Advances in Information Retrieval. Berlin: Springer Berlin Heidelberg, 2011: 338-349.

[24] IWATA T, WATANABE S, YAMADA T, et al. Topic Tracking Model for Analyzing Consumer Purchase Behavior[C]// IJCAI. 2009: 1427-1432.

[25] SASAKI K, YOSHIKAWA T, FURUHASHI T. Twitter-TTM: An efficient online topic modeling for Twitter considering dynamics of user interests and topic trends[C]// 15th International Symposium on Soft Computing and Intelligent Systems (SCIS), 2014 Joint 7th International Conference on and Advanced Intelligent Systems (ISIS). IEEE, 2014: 440-445.

[26] ZHANG C Y, SUN J L, DING Y Q. Topic Mining for Microblog Based on MB-LDA Model[J]. Journal of Computer Research and Development, 2011, 48(10): 1795-1802. (in chinese)
张晨逸，孙建伶，丁轶群. 基于 MB-LDA 模型的微博主题挖掘[J]. 计算机研究与发展，2011，48(10)：1795-1802.

[27] LI J, YIN J, LIU S P, et al. Microblog Topic Mining Based on Hashtag[J]. Computer Engineering, 2015, 41(4): 30-35. (in Chinese)
李敬，印鉴，刘少鹏，等. 基于话题标签的微博主题挖掘[J]. 计算机工程，2015，41(4)：30-35.

[28] 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): 3491-3496. (in Chinese)
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)：3491-3496.