第 40 卷 第 5 期
Vol.40 No.5

计 算 机 工 程
Computer Engineering

2014 年 5 月
May 2014

• 移动社交专题 •

文章编号：1000-3428(2014)05-0001-06 文献标识码：A 中图分类号：TP311.13

LDA 模型在微博用户推荐中的应用

邸 亮，杜永萍

(北京工业大学计算机科学与技术学院，北京 100124)

摘 要：潜在狄利克雷分配(LDA)主题模型可用于识别大规模文档集中潜藏的主题信息，但是对于微博短文本的应用效果并不理想。为此，提出一种基于 LDA 的微博用户模型，将微博基于用户进行划分，合并每个用户发布的微博以代表用户，标准的文档-主题-词的三层 LDA 模型变为用户-主题-词的用户模型，利用该模型进行用户推荐。在真实微博数据集上的实验结果表明，与传统的向量空间模型方法相比，采用该方法进行用户推荐具有更好的效果，在选择合适的主题数情况下，其准确率提高近 10%。

关键词：主题模型；潜在狄利克雷分配；微博；用户模型；兴趣分析；用户推荐

Application of LDA Model in Microblog User Recommendation

DI Liang, DU Yong-ping

(Institute of Computer Science and Technology, Beijing University of Technology, Beijing 100124, China)

【Abstract】 Latent Dirichlet Allocation(LDA) model can be used for identifying topic information from large-scale document set, but the effect is not ideal for short text such as microblog. This paper proposes a microblog user model based on LDA, which divides microblog based on user and represents each user with their posted microbolgs. Thus, the standard three layers in LDA model by document-topic-word becomes a user model by user-topic-word. The model is applied to user recommendation. Experiment on real data set shows that the new provided method has a better effect. With a proper topic number, the performance is improved by nearly 10%.

【Key words】 topic model; Latent Dirichlet Allocation(LDA); microblog; user model; interest analysis; user recommendation

DOI: 10.3969/j.issn.1000-3428.2014.05.001

1 概述

传统的主题挖掘是采用文本聚类的算法[1]，通过向量空间模型(Vector Space Model, VSM)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，如基于划分的算法(如 K-means 算法)、基于层次的算法(如自顶向下和自底向上算法)、基于密度的算法等[2]，实现文本聚类。聚类结果可以近似认为满足同一个主题。但是，这种基于聚类的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。

LSA(Latent Semantic Analysis)是文献[3]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，通过奇异值分解等数学手段，使得这种隐含的相关性能够被很好地挖掘出来。研究显示[4]，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，即将表面信息转化为深层次的抽象[5]。

PLSA(Probabilistic Latent Semantic Analysis)是文献[6]在研究 LSA 的基础上提出的基于最大似然法和产生式模型的概率模型。PLSA 沿用了 LSA 的降维思想：在常用的文本表达方式(tf-idf)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。

潜在狄利克雷分配(Latent Dirichlet Allocation, 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现象，并且很难对训练集以外的文档

基金项目：国家科技支撑计划基金资助项目(2013BAH21B00)；北京市自然科学基金资助项目(4123091)；北京市属高等学校人才强教深化计划基金资助项目“中青年骨干人才培养计划”(PHR20110815)。

作者简介：邸 亮(1988－)，男，硕士研究生，主研方向：自然语言处理；杜永萍，副教授。

收稿日期：2013-09-22 **修回日期：**2013-12-05 **E-mail：**dltt67@163.com

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-单词三层的贝叶斯模型[7]，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。目前，LDA 模型已经成为了主题建模中的一个标准，在多个领域中都有应用，特别是在社会网络和社会媒体研究领域最为常见[8]，具有很好的研究与应用前景。在微博主题挖掘中具有很大的潜力[9-10]，通过对其进行改进，可以很好地应用于社交网络应用中。

本文在 LDA 主题模型的基础上，通过分析微博用户的特点，给出了用以表示用户主题的模型，并提出一种基于该模型的用户推荐方法。

2 LDA 主题模型

LDA 模型是一个层次贝叶斯模型[11]，它有如下 3 层：

(1)单词层：单词集 $\boldsymbol{V}=\{w_1,w_2,\cdots,w_V\}$ 是从语料库中提取出来的去除停用词后的所有单词集合。

(2)主题层：主题集 $\boldsymbol{\phi}=\{z_1,z_2,\cdots,z_k\}$ 中的每一个主题 z_i 都是一个基于单词集 $\boldsymbol{V}$ 的概率多项分布，可以被表示成向量 $\boldsymbol{\varphi}_k=<p_{k,1},p_{k,2},\cdots,p_{k,v}>$，其中，$p_{k,j}$ 表示单词 w_j 在主题 z_k 中的生成概率。

(3)文档层：对于单词层，采用了词袋方法。每一篇文档被表示成一个词频向量 $\boldsymbol{d}_i=<tf_{i,1},tf_{i,2},\cdots,tf_{i,V}>$，其中，$tf_{i,j}$ 表示单词 j 在文档 i 中出现的次数；就主题层而言，文档集可以表示成 $\boldsymbol{\theta}=<\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\cdots,\boldsymbol{\theta}_D>$，其中每一个向量 $\boldsymbol{\theta}_d=<p_{d,1},p_{d,2},\cdots,p_{d,K}>$ 表示了一个文档的主题分布，$p_{d,z}$ 是主题 z 在该文档 d 中的生成概率。

其图模型表示如图 1 所示。LDA 模型采用 Dirichlet 分布作为概率主题模型中多项分布的先验分布。其中，$\boldsymbol{D}$ 为整个文档集；$\boldsymbol{N}_d$ 为文档 $\boldsymbol{d}$ 的单词集；α 和 β 分别是文档-主题概率分布 θ 和主题-单词概率分布 ϕ 的先验知识。

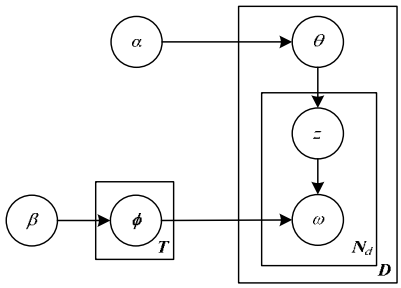

图 1 LDA 图模型

3 基于 LDA 模型的微博用户推荐

3.1 基于 LDA 模型的微博用户模型

标准的 LDA 模型是基于文档-主题-词的一个三层贝叶斯模型[11]。在构建用户的兴趣模型时，用户的兴趣可以被定义为用户对各个主题的喜好程度。因此，主题模型下用户-主题生成概率多项分布表示了用户的兴趣。

使用主题模型构建基于内容的微博用户兴趣模型时，需要将一个用户下的所有微博合并成一个文档进行主题生成，从而得到用户生成主题的概率多项分布，即用户的兴趣模型。该兴趣模型的用户层就对应到了 LDA 模型中的文档层，即将文档-主题-词的三层关系变为了用户-主题-词的关系，其矩阵表示如图 2 和图 3 所示。

$$\underset{\text{词语}}{\begin{matrix}w_1\\w_2\\\vdots\\w_n\end{matrix}}\overset{\text{文档}}{\begin{matrix}d_1&d_2&\cdots&d_m\end{matrix}}\begin{Bmatrix}&&&\\&&&\\&\boldsymbol{C}&&\\&&&\\&&&\end{Bmatrix}=\underset{\text{词语}}{\begin{matrix}w_1\\w_2\\\vdots\\w_n\end{matrix}}\overset{\text{主题}}{\begin{matrix}z_1&z_2&\cdots&z_k\end{matrix}}\begin{Bmatrix}&&&\\&&&\\&\boldsymbol{\Phi}&&\\&&&\\&&&\end{Bmatrix}\times\underset{\text{主题}}{\begin{matrix}z_1\\z_2\\\vdots\\z_k\end{matrix}}\overset{\text{文档}}{\begin{matrix}d_1&d_2&\cdots&d_m\end{matrix}}\begin{Bmatrix}&&&\\&&&\\&\boldsymbol{\Theta}&&\\&&&\\&&&\end{Bmatrix}$$

图 2 标准 LDA 模型的矩阵示意图

$$\underset{\text{词语}}{\begin{matrix}w_1\\w_2\\\vdots\\w_n\end{matrix}}\overset{\text{用户}}{\begin{matrix}u_1&u_2&\cdots&u_m\end{matrix}}\begin{Bmatrix}&&&\\&&&\\&\boldsymbol{U}&&\\&&&\\&&&\end{Bmatrix}=\underset{\text{词语}}{\begin{matrix}w_1\\w_2\\\vdots\\w_n\end{matrix}}\overset{\text{主题}}{\begin{matrix}z_1&z_2&\cdots&z_k\end{matrix}}\begin{Bmatrix}&&&\\&&&\\&\boldsymbol{W}&&\\&&&\\&&&\end{Bmatrix}\times\underset{\text{主题}}{\begin{matrix}z_1\\z_2\\\vdots\\z_k\end{matrix}}\overset{\text{用户}}{\begin{matrix}u_1&u_2&\cdots&u_m\end{matrix}}\begin{Bmatrix}&&&\\&&&\\&\boldsymbol{K}&&\\&&&\\&&&\end{Bmatrix}$$

图 3 基于 LDA 的微博用户模型的矩阵示意图

在用户层中，对于用户集合 $\boldsymbol{U}=\{u_1,u_2,\cdots,u_m\}$，其中的每一个用户 u_i，都可以由该用户发布的所有微博得到一个词频向量 $\boldsymbol{f}_{u_i}=<tf_{i,1},tf_{i,2},\cdots,tf_{i,V}>$。从主题层面而言，用户 u_i 可以被表示成向量 $\boldsymbol{\theta}_{u_i}=\{p_{u_i,1},p_{u_i,2},\cdots,p_{u_i,k}\}$，其中，$p_{u_i,z}$ 表示主题 z 在用户 u_i 中的生成概率，用它来表示用户 u_i 对主题 z 的喜好程度。从而，用户层构成了用户与主题的生成关系，生成主题用户模型，其矩阵表示如图 4 所示。

$$\begin{array}{c} \quad z_1 \quad z_2 \quad \cdots \quad z_k \\ u_1 \\ u_2 \\ \vdots \\ u_m \end{array} \begin{pmatrix} p_{u_1,1}, p_{u_1,2}, \cdots, p_{u_1,k} \\ p_{u_2,1}, p_{u_2,2}, \cdots, p_{u_2,k} \\ \vdots \\ p_{u_m,1}, p_{u_m,2}, \cdots, p_{u_m,k} \end{pmatrix}$$

图 4　用户主题矩阵

3.2　用户相似度计算

KL(Kullback Leibler)散度，俗称 KL 距离[12]，常用来衡量 2 个概率分布的距离，其计算公式如下：

$$D_{KL}(\boldsymbol{P} \| \boldsymbol{Q}) = \sum_i \boldsymbol{P}(i) \ln \frac{\boldsymbol{P}(i)}{\boldsymbol{Q}(i)} \tag{1}$$

KL 散度是不对称的，即 $D_{KL}(\boldsymbol{P} \| \boldsymbol{Q}) \neq D_{KL}(\boldsymbol{Q} \| \boldsymbol{P})$，可以将其转换为对称的，如下式：

$$D(\boldsymbol{P}, \boldsymbol{Q}) = [D_{KL}(\boldsymbol{P} \| \boldsymbol{Q}) + D_{KL}(\boldsymbol{Q} \| \boldsymbol{P})] / 2 \tag{2}$$

在基于 LDA 的用户主题模型中，由主题的概率分布来表示用户的兴趣，如图 4 用户主题矩阵所示。因此，用户间的相似程度可以由用户主题分布间的 KL 距离来表示，用户相似度计算如下所示：

$$S_{ij} = \frac{1}{D(\boldsymbol{U}_i, \boldsymbol{U}_j)} = \frac{2}{[D_{KL}(\boldsymbol{U}_i \| \boldsymbol{U}_j) + D_{KL}(\boldsymbol{U}_j \| \boldsymbol{U}_i)]} \tag{3}$$

其中，S_{ij} 为用户 u_i 和 u_j 的相似度；$\boldsymbol{U}_i$ 和 $\boldsymbol{U}_j$ 分别是它们的主题概率分布。该值越大，则两用户越相似。

3.3　用户推荐

假设同一个领域中的用户为兴趣相近的用户，且他们的微博也主要是围绕自己感兴趣的话题来发布。

$\boldsymbol{U}$ 为用户集合，对用户 u_i 和用户子集 $\boldsymbol{U}_i$，其中，$u_i \in \boldsymbol{U}$，且 $\boldsymbol{U}_i = \boldsymbol{U} - u_i$。按照式(3)，对用户集合 $\boldsymbol{U}_i$ 中的每个用户分别与 u_i 计算相似度，然后对 $\boldsymbol{U}_i$ 中的所有用户按照相似度值进行升序排列，这样排在前面的用户就和用户 u_i 更相似，更有理由推荐给用户 u_i。

提取前 t 个用户作为推荐给用户 u_i 的推荐列表，$\boldsymbol{U}_{t_i} = \{u_1, u_2, \cdots, u_j, \cdots, u_t\}$。对推荐集合 $\boldsymbol{U}_{t_i}$ 中的每个用户 u_j，分别判断其是否与用户 u_i 属于同一领域，若属于同一领域，则认为将 u_j 推荐给用户 u_i 是正确的。用户 u_i 的推荐准确率计算公式如下：

$$Accuracy(u_i) = \frac{\sum_{u_j \in \boldsymbol{U}_{t_i}} f(u_i, u_j)}{k}，\ t \leq N_i - 1 \tag{4}$$

$$f(u_i, u_j) = \begin{cases} 1 & u_i, u_j \text{属于同一领域} \\ 0 & u_i, u_j \text{不属于同一领域} \end{cases} \tag{5}$$

其中，$t \leq N_i - 1$，N_i 为用户 u_i 所属领域下的用户数，t 的取值不超过该领域下的用户总数减 1(除去用户 u_i 自身)。

某领域 p 下用户的推荐准确率计算公式如下：

$$Accuracy(p) = \frac{\sum_{i=1}^{N_p} Accuracy(u_i)}{N_p} \tag{6}$$

其中，N_p 为领域 p 下的用户总数。

在系统中，所有用户的推荐平均准确率计算公式如下：

$$\overline{Accuracy} = \frac{\sum_{i=1}^{N} Accuracy(u_i)}{N} \tag{7}$$

其中，N 为用户总数。

3.4　用户推荐系统结构

基于上文介绍的用户兴趣模型，设计了微博用户推荐系统，主要由 3 个部分组成：

(1)数据采集层，负责微博数据的采集及预处理，预处理包括对部分字数过少微博的过滤。

(2)数据处理层，对过滤后的微博数据做进一步处理，包括分词、去停用词、词性过滤等，生成用户的词语向量，从而得到整个用户集合的向量表示，利用 LDA 用户模型进行求解，从而进行主题挖掘和用户推荐。

(3)数据展现层，展现数据处理层生成的结果，包括模型生成的主题的展示、用户推荐的关联图等。

系统结构如图 5 所示。

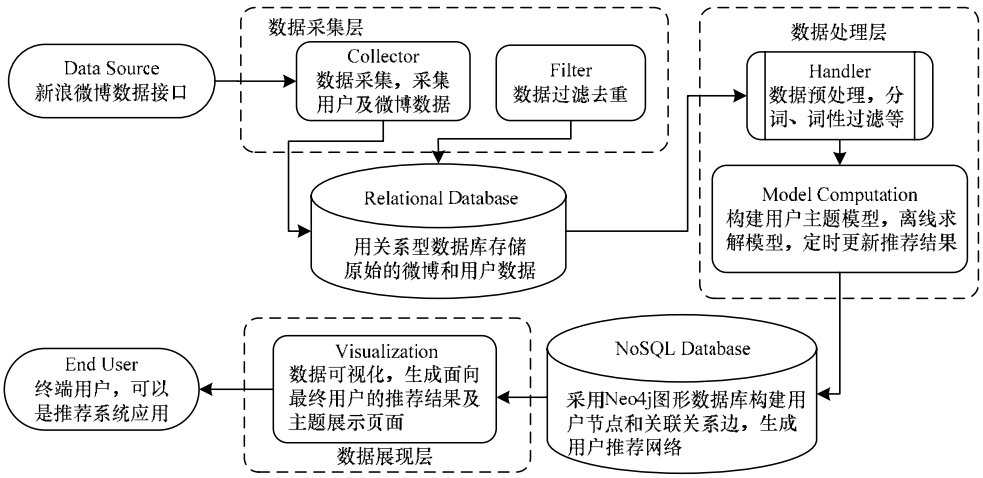

图 5　用户推荐系统结构

在图 5 中涉及到的关键技术主要有：

(1)数据采集器使用开源的 Java 工具包 HttpClient 实现。调用新浪微博 API 后，获取到 json 格式的数据，需要将其解析为数据对象，然后存入数据库。

(2)微博及用户数据采用关系型数据库来保存。这里使用 MySQL，因为其体积小、速度快，并且是开源的。

(3)数据处理过程中用到了哈工大的 IRLAS 分词器，对微博进行分词和词性标注。

(4)构造出主题模型后，将用户推荐结果存入 NoSQL 数据库，这里使用 Neo4j，它是一个用 Java 实现、完全兼容 ACID 的图形数据库，数据以一种针对图形网络进行过优化的格式保存在磁盘上，它的内核是一种极快的图形引擎，具有数据库产品期望的所有特性。用 Neo4j 存储用户推荐结果可以方便快速地实现前台的展示。

(5)可视化主要通过 js 及其第三方开源库来实现，例如 D3 js 库可以实现主题关键词的标签云展示及用户推荐的关联散点图等。

3.5 算法流程

基于 LDA 模型的微博用户推荐算法如下：

(1)建立用户模型：将用户的所有微博合并到一起，微博数据已经经过了分词处理，得到代表每个用户的微博单词词频向量 $\boldsymbol{f}_u$。对模型进行求解，得到每个用户的主题概率分布，如图 4 所示。

(2)用户相似度计算：借助于概率分布之间的 KL 散度计算方法，用户之间的相似度使用式(3)来计算，该值越大则表示用户间的主题概率分布越相似，也即用户间的兴趣越相似，双方可以相互作为被推荐给对方的候选用户。

(3)用户推荐：假设同一个领域中的用户为兴趣相近的用户，根据用户相似度获取用户的推荐列表，取前 t 个用户作为推荐用户，利用式(4)~式(7)计算推荐准确率。

4 实验结果与分析

4.1 数据采集与预处理

实验利用新浪微博 API 采集用户数据和微博数据。主要用到 2 个接口：获取系统推荐的热门用户列表接口和获取单个用户微博列表的接口。

根据推荐用户接口抓取来自不同领域的认证用户数据，获取了 8 个比较常见的领域，分别是科技、体育、房产、动漫、娱乐、健康、汽车和媒体。此外，利用用户微博列表接口采集每个用户的最新微博，最多不超过 300 条。

由于微博数据来自于互联网，噪声大，需要做一定的预处理，主要有以下 4 个步骤：

(1)将回复数和转发数低于 10 的微博去除。

(2)根据用户实际有效的微博数量，从每个领域中各选取 80 个用户。选取的过程会过滤掉有效微博数量小于 10 条的用户，最终实验数据集的总用户数为 640 个。

(3)去掉微博数据中特有的一些对主题挖掘无用的特征，如表情符号、@目标、分享目标以及 URL 网址等。

(4)对微博数据进行分词，过滤掉停用词，根据词性标注保留对主题挖掘提供有用的信息的名词、动词。

最终用于实验的数据组成如表 1 所示。

表 1 实验数据分布

领域	用户数量	有效微博数量
科技领域	80	22 399
体育领域	80	11 434
房产领域	80	22 724
动漫领域	80	27 317
娱乐领域	80	21 913
健康领域	80	20 819
汽车领域	80	17 013
媒体领域	80	23 994
总计	640	167 613

4.2 实验参数设置与对比实验

LDA 模型的求解过程使用 Gibbs 抽样方法，模型参数值根据文献[11]取经验值：其中，α=50/T(T 为主题数)，β=0.01。主题的个数取经验值进行对比实验，由于用户来自于 8 个领域，实验中主题数设置为 8~15。分词器采用哈工大 IRLAS 分词器，使用通用停用词词典，共 1 241 条停用词项。

为了进一步对比实验效果，把本文算法与下面 2 个算法进行比较：

(1)基于向量空间模型(VSM)的算法

使用传统的 VSM 方法建立用户模型，同样对于用户集 $\boldsymbol{U}=\{u_1,u_2,\cdots,u_m\}$，将用户 u_i 的所有微博数据进行预处理后得到其单词权重向量 $\boldsymbol{U}_i=<w_{i,1},w_{i,2},\cdots,w_{i,V}>$，其中，$w_{i,j}$ 表示单词 j 在用户 u_i 的微博数据中的权重。这里的权重计算采用 TF-IDF 值。用户间相似度的计算采用常规的向量夹角的余弦值来计算：

$$Sim(u_i,u_j)=\frac{\boldsymbol{U}_i\cdot\boldsymbol{U}_j}{|\boldsymbol{U}_i^2\|\boldsymbol{U}_j^2|} \tag{8}$$

(2)基于隐马尔科夫模型(HMM)的算法

应用文献[13]中介绍的方法。使用 HMM 建立用户的模型，$\boldsymbol{\lambda}=(A,B,\pi,N,M)$，然后使用 KL 散度计算用户间的相似度，计算公式为：

$$D_{KL}(\boldsymbol{\lambda}_m,\boldsymbol{\lambda}_n)=\frac{1}{V}(\ln P(w_s\mid\boldsymbol{\lambda}_m)-\ln P(w_s\mid\boldsymbol{\lambda}_n))$$

以上 2 种算法的用户推荐准确率的计算方法和 LDA 用户模型的计算方法相同，不再赘述。

4.3 评价结果

4.3.1 基于 Perplexity 指标的评价结果

Perplexity[9]是一种评估语言模型生成性能的标准测量指标。Perplexity 值表示模型生成测试集中新文本的似然估计，它用来衡量模型对新文本的预测能力。Perplexity 值越

小，似然估计就越高，也就表示模型的生成性能越好。其计算公式如下：

$$Perplexity(\boldsymbol{U}_{\text{test}}) = \exp\{-\frac{\sum_{i=1}^{N}\ln p(\boldsymbol{w}_{u_i})}{\sum_{i=1}^{N} N_{u_i}}\} \tag{9}$$

其中，$\boldsymbol{U}_{\text{test}}$ 为测试集用户；N 为测试集用户总数；$\boldsymbol{w}_{u_i}$ 为用户 u_i 的微博所包含的单词集合；$p(\boldsymbol{w}_{u_i})$ 是用户 u_i 的微博单词集合在用户模型下的生成概率；N_{u_i} 为用户 u_i 微博集合的单词总数。实验中选取了数据集的 10%作为测试集。

实验结果如图 6 所示。

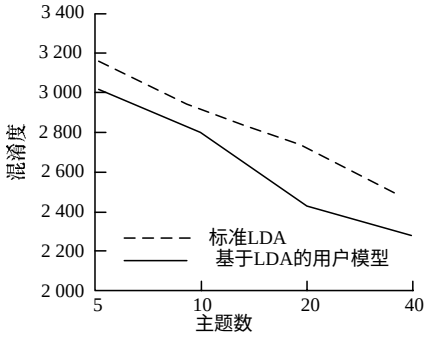

图 6　用户兴趣模型的 Perplexity 评价结果

从图 6 中的数据可以看出，基于 LDA 的用户模型的生成能力要优于标准 LDA，这说明将同一用户的微博合并为一条文本的方式是有效的。

4.3.2　主题分布

选取一些有代表性的主题分布生成的标签云图，如图 7 所示，可以很明显地看出，这些主题分布分别代表了科技、体育、房产、动漫、娱乐、健康、汽车、媒体相关的主题。

图 7　主题分布词云图

4.3.3　用户推荐质量

用户推荐质量的衡量需要从实际的应用效果入手，由于该模型可以对具有相似兴趣的用户进行推荐，这里使用上述介绍的用户推荐准确率来衡量模型的质量。LDA 用户模型和 VSM 方法在各领域下的准确率对比结果如表 2~表 5 所示，分别对应式(4)中 t 取 10，20，40，79 时的结果。

表 2　t=10 时的实验结果

领域	LDA 用户模型(主题数为 K)									VSM 模型	HMM 模型
	K=8	K=9	K=10	K=11	K=12	K=13	K=14	K=15	K=20		
体育领域	0.740	0.750	0.766	0.670	0.777	0.736	0.820	0.747	0.811	0.691	0.712
房产领域	0.240	0.235	0.280	0.426	0.410	0.469	0.490	0.485	0.490	0.320	0.471
动漫领域	0.540	0.428	0.565	0.615	0.620	0.620	0.600	0.636	0.701	0.500	0.633
娱乐领域	0.630	0.585	0.606	0.630	0.673	0.659	0.673	0.645	0.681	0.510	0.620
科技领域	0.480	0.500	0.540	0.670	0.650	0.664	0.687	0.685	0.727	0.417	0.661
汽车领域	0.300	0.340	0.385	0.410	0.344	0.366	0.500	0.482	0.527	0.400	0.413
健康领域	0.533	0.520	0.568	0.540	0.560	0.576	0.563	0.534	0.571	0.625	0.540
媒体领域	0.625	0.674	0.550	0.690	0.696	0.680	0.661	0.696	0.666	0.585	0.691
平均	0.511	0.504	0.532	0.581	0.591	0.596	0.624	0.614	0.647	0.506	0.592

表 3　t=20 时的实验结果

领域	LDA 用户模型(主题数为 K)									VSM 模型	HMM 模型
	K=8	K=9	K=10	K=11	K=12	K=13	K=14	K=15	K=20		
体育领域	0.708	0.711	0.686	0.614	0.710	0.656	0.760	0.676	0.736	0.562	0.632
房产领域	0.245	0.236	0.283	0.420	0.387	0.446	0.451	0.428	0.441	0.286	0.432
动漫领域	0.503	0.420	0.550	0.591	0.591	0.597	0.571	0.626	0.642	0.457	0.552
娱乐领域	0.591	0.542	0.576	0.595	0.632	0.640	0.633	0.630	0.634	0.447	0.591
科技领域	0.464	0.500	0.533	0.656	0.636	0.648	0.685	0.683	0.706	0.400	0.583
汽车领域	0.267	0.333	0.353	0.371	0.309	0.348	0.469	0.466	0.500	0.348	0.391
健康领域	0.484	0.486	0.531	0.515	0.514	0.540	0.514	0.500	0.517	0.569	0.523
媒体领域	0.534	0.621	0.487	0.534	0.660	0.624	0.569	0.610	0.563	0.491	0.587
平均	0.475	0.481	0.500	0.537	0.554	0.562	0.581	0.577	0.592	0.445	0.536

表 4 t=40 时的实验结果

领域	LDA 用户模型(主题数为 K)									VSM 模型	HMM 模型
	K=8	K=9	K=10	K=11	K=12	K=13	K=14	K=15	K=20		
体育领域	0.613	0.608	0.571	0.525	0.609	0.549	0.667	0.623	0.647	0.423	0.580
房产领域	0.233	0.244	0.264	0.367	0.349	0.382	0.372	0.369	0.368	0.250	0.380
动漫领域	0.450	0.385	0.495	0.556	0.550	0.561	0.536	0.591	0.558	0.406	0.545
娱乐领域	0.558	0.525	0.527	0.535	0.567	0.567	0.571	0.558	0.563	0.321	0.520
科技领域	0.453	0.488	0.498	0.612	0.590	0.617	0.658	0.643	0.674	0.359	0.500
汽车领域	0.250	0.313	0.319	0.349	0.284	0.313	0.425	0.418	0.447	0.300	0.333
健康领域	0.450	0.425	0.479	0.473	0.455	0.473	0.456	0.445	0.416	0.446	0.464
媒体领域	0.441	0.515	0.400	0.522	0.514	0.496	0.467	0.470	0.450	0.409	0.407
平均	0.431	0.438	0.444	0.492	0.490	0.495	0.519	0.515	0.515	0.364	0.466

表 5 t=79 时的实验结果

领域	LDA 用户模型(主题数为 K)									VSM 模型	HMM 模型
	K=8	K=9	K=10	K=11	K=12	K=13	K=14	K=15	K=20		
体育领域	0.489	0.494	0.456	0.447	0.517	0.463	0.545	0.503	0.507	0.427	0.463
房产领域	0.229	0.238	0.243	0.293	0.287	0.313	0.319	0.308	0.300	0.253	0.260
动漫领域	0.376	0.337	0.416	0.472	0.458	0.463	0.447	0.463	0.442	0.408	0.454
娱乐领域	0.403	0.395	0.399	0.400	0.415	0.417	0.420	0.412	0.409	0.324	0.391
科技领域	0.426	0.438	0.440	0.530	0.519	0.539	0.557	0.547	0.556	0.360	0.482
汽车领域	0.233	0.280	0.286	0.317	0.250	0.274	0.332	0.327	0.352	0.300	0.291
健康领域	0.334	0.313	0.344	0.328	0.323	0.332	0.325	0.307	0.295	0.449	0.400
媒体领域	0.289	0.325	0.260	0.327	0.336	0.307	0.296	0.292	0.285	0.408	0.310
平均	0.347	0.352	0.355	0.389	0.388	0.388	0.405	0.395	0.393	0.366	0.381

分析以上实验结果得出结论：

(1)推荐性能与主题数相关。随着主题数的增加，推荐效果逐渐变好，在主题数为 14 时，推荐效果最好，当主题数进一步增加时，效果基本保持稳定甚至略微有所回落。主题数越大，模型的计算量也越大，耗时越久，综合可虑，在主题数取 14 时，无论是推荐效果还是计算效率都有着不错的结果。对比 VSM 模型的实验结果后还可以看出，当主题数大于 10 的情况下，基于 LDA 的用户兴趣模型的效果均比传统的 VSM 有所提高。而对比 HMM 模型的实验结果可以看出，当主题数达到 12 时，基于 LDA 的用户兴趣模型的效果和 HMM 模型相当，在主题数大于 14 的情况下，效果明显好于 HMM 模型。

(2)推荐性能在不同领域下有着较明显的差别。LDA 用户兴趣模型对体育领域和科技领域的用户推荐效果较好，尤其是体育领域，K 取 10 时其准确率甚至达到了 82%，远好于其他领域。房产和汽车领域的效果略微偏差，分析这些领域用户的微博，发现这可能是由于这些领域用户发布的微博比较宽泛，涉及的内容和主题比较繁杂，对主题挖掘的干扰比较大；而体育领域和科技领域的用户发布的微博则相对更具有明确的主题，领域凝聚力更强，实用性更高，因此更有挖掘主题的价值。如何减少这类微博对用户推荐的干扰，是今后的工作重点。

5 结束语

本文针对微博数据这种短文本，结合 LDA 模型的文档-主题-词分层模型的特点，用微博数据的集合来代表用户，进而提出了用户-主题-词的用户兴趣模型，不仅能有效挖掘用户所关注的主题，并可进行用户推荐等社交网络应用。在今后的研究工作中将继续优化微博用户兴趣模型的效果和效率，减少无意义微博对主题挖掘的干扰，以适应于各种不同的领域，尝试结合更多的社交网络特征，并实现实时的微博数据处理。

参考文献

[1] Kang J H, Lerman K, Plangprasopchok A. Analyzing Micro-blogs with Affinity Propagation[C]//Proc. of the 1st Workshop on Social Media Analytics. New York, USA: ACM Press, 2010: 67-70.

[2] Xu Rui, Wunsch D. Survey of Clustering Algorithms[J]. IEEE Trans. on Neural Networks, 2005, 16(3): 645-678.

[3] Deerwester S, Dumais S, Landauer T, et al. Latent Semantic Analysis for Multiple-type Interrelated Data Objects[C]//Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2006: 236-243.

(下转第 11 页)