Software Engineering and Applications 软件工程与应用, 2017, 6(5), 145-153
Published Online October 2017 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2017.65016

Hans汉斯

Hotspot Topics Detection from WeChat Library Based on Model LDA

Jing Xun

Xi'an Technological University, Xi'an Shaanxi
Email: xunjing311416@163.com

Received: Oct. 5th, 2017; accepted: Oct. 17th, 2017; published: Oct. 24th, 2017

Abstract

In order to make the library staff relieve from a large amount of redundant information and real-time understanding of the needs of teachers and students, for WeChat library, in the paper, the method of hotspot topic detection based on model Latent Dirichlet Allocation (LDA) was proposed. The method first merged the characteristic words by constructing the professional dictionary in the library field, and then all the texts of WeChat were described by model LDA. Finally, the similarity between texts was calculated by topic similarity, and then the Single-Pass clustering algorithm was used to cluster WeChat data and found hotspot topics. The experimental results show that this method can effectively identify hotspot topics, and achieve good results in precision, recall and F-measure.

Keywords

Model Latent Dirichlet Allocation (LDA), Topic Similarity, Single-Pass Clustering Algorithm, Topic Detection

基于LDA模型的微信图书馆热点话题检测

荀　静

西安工业大学，陕西　西安
Email: xunjing311416@163.com

收稿日期：2017年10月5日；录用日期：2017年10月17日；发布日期：2017年10月24日

摘　要

为使图书馆工作人员免受大量冗余信息的困扰，实时了解广大师生的需求及关注热点，面向微信图书馆，本文给出一种基于LDA模型的微信热点话题检测方法。该方法首先通过构建图书馆领域专业词典合并特

文章引用：荀静. 基于 LDA 模型的微信图书馆热点话题检测[J]. 软件工程与应用, 2017, 6(5): 145-153.
DOI: 10.12677/sea.2017.65016

征词，其次应用LDA模型表示微信文本信息，最后采用主题相似度计算文本间的相似度，进而利用Single-Pass聚类算法识别热点话题。实验结果表明，该方法能够有效地对微信图书馆上的数据进行话题检测，在准确率、召回率和F_1值上均有不错的效果。

关键词

LDA模型，主题相似度，Single-Pass聚类算法，话题检测

1. 引言

近年来，随着互联网社交网络的兴起，微博、twitter 和微信等社交品台受到人们的广泛关注。其中，微信公众平台是腾讯公司主要面向名人、政府、媒体、企业等机构推出的合作推广业务，旨在为用户提供服务和信息。用户通过一键操作的形式可以便捷地使用微信公众平台的资源和服务，给用户带来了极大的便利。

高校图书馆一直将“以人为本”作为服务理念，致力于为全校师生提供更全面更方便的阅读服务系统，而微信公众平台正好能够落实这一理念，为图书馆的发展与提升提供充分的保障。高校图书馆现已陆续开通微信公众号，既符合时下大学生的生活方式和阅读习惯，也能够让学校师生随时随地的使用图书馆数据资源。用户每天都会在微信公众平台上发布大量的评论信息和反馈信息，如何获取这些信息中的有用价值，把握用户的关注热点、兴趣偏好和情感倾向，需要我们进一步的挖掘。

微信中的数据以短文本的形式存在，如果采用传统的向量空间模型建模，容易造成高维性与稀疏性，从而影响话题聚类效果。鉴于向量空间模型存在的缺点，研究学者引入以 LDA 模型为代表的主题模型对短文本进行处理。LDA 主题模型是一种无监督的机器学习算法，能够将每一条文本数据转化为语义空间的向量，已在微博、微信和网络论坛等短文本研究领域得到广泛应用。

Quan 等[1]利用主题模型表示文本，然后通过查找两篇文本的可区分词集，在主题的基础上度量文本间的相似性。张志飞等[2]在 Quan 研究的基础上利用 LDA 模型生成的主题，进一步解决短文本上下文依赖性的问题。在社交平台热点话题识别方面，孙励[3]使用 LDA 模型挖掘微博数据的主题信息进而发现热点话题。刘红兵等[4]结合微博短文本的相关特点，提出一种基于 LDA 模型和多层聚类的微博话题检测方法。汪进洋[5]将中文词性标注和 LDA 主题模型两种方法用于微博话题的检测，并使用增量聚类方法确定微博话题个数和微博聚类。余传明等[6]利用潜在狄利克雷分布模型与自然语言处理技术相结合的方法挖掘用户评论数据，获取评论热点及相应的热点词语。

在已有研究的基础上，本文提出一种基于 LDA 模型的微信图书馆热点话题检测方法。该方法利用 LDA 模型挖掘用户时间窗口内的数据，解决微信图书馆数据的稀疏问题，然后使用主题相似度解决短文本文本之间相似度的计算问题，最后通过 Single-Pass 聚类算法进行微信图书馆话题检测。实验结果表明，该方法可以有效地识别微信图书馆中的热点话题。

2. 相关工作

2.1. LDA 模型

微信图书馆中的数据具有更新快、文本短等特性，如果使用 LDA 模型直接表示用户的微信文本信息，

得到的主题会比较稀疏，从而影响后续话题聚类效果。用户在短时间间隔内通常是关注单一话题的，因此本文引入时间窗口，把用户在时间窗口内发布的所有微信信息合并成一个文本信息，然后将合并后的文本信息应用到 LDA 模型中。

LDA 主题模型是由 Blei 等[7]人提出的一种统计机器学习模型，已被广泛用于文档、图像等信息的建模。LDA 是一个“文本－主题－词”的三层贝叶斯模型，能够对大规模文本集合进行降维处理，识别大规模文本集合中隐藏的主题信息。LDA 模型对文本进行一个简略的叙述，保留本质的统计信息，从而可以快速有效地处理大量的文本集合。其把每一篇文本看作一个词频向量，从而将计算机难以识别的文本信息转变成易于建模的数字信息。每一篇文本代表一些主题组成的一个概率分布，同时，每一个主题代表大量词汇标记组成的一个概率分布。

LDA 模型的表示如图 1 所示。

图 1 代表的联合概率模型为：

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \tag{1}$$

对公式 1 计算边缘概率，可以得到：

$$p(D|\alpha, \beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta)\right)d_{\theta_d} \tag{2}$$

其中，d 表示文本，z 表示潜在主题；α 和 β 是 LDA 模型的 Dirichlet 先验分布，α 是文本集中含有的所有主题分布的先验，β 是所有主题中含有的全部词汇分布的先验；θ_d 代表文本 d 中包含的所有主题的多项式分布，φ_z 代表主题 z 中包含的所有词语的多项式分布。

本文采用 Gibbs 抽样算法[8]估计参数 θ 和 φ，舍弃词汇标记，以 t 表示唯一性的词，则对于每个词汇来说，θ_{mk} 和 φ_{kt} 的值可估计如下：

$$\theta_{mk} = \frac{n_{m,\neg i}^{(k)} + \alpha}{\sum_{j=1}^{K} n_{m,\neg i}^{(j)} + K\alpha} \tag{3}$$

$$\varphi_{kt} = \frac{n_{k,\neg i}^{(t)} + \beta}{\sum_{t=1}^{N} n_{k,\neg i}^{(t)} + N\beta} \tag{4}$$

其中，θ 代表在文本 m 中从主题 k 抽取词汇的概率预测，φ 代表从主题 k 中抽到词汇标记为 t 的词汇的概率预测。

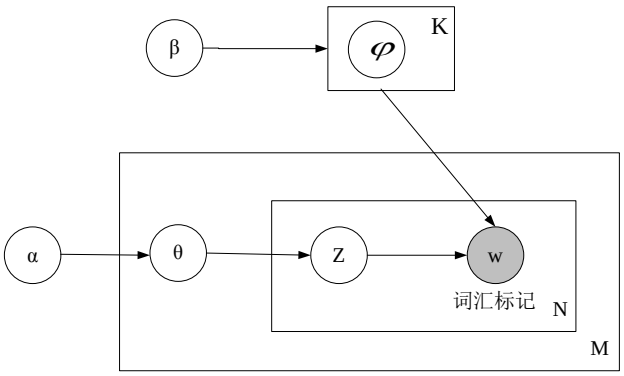

Figure 1. LDA model generates text process
图 1. LDA 模型生成文本过程

2.2. 主题相似度

由于微信数据具有词汇量少、文本短和文本数量多等特点，传统的余弦相似度算法不能很好的反映两个文本间的相似程度。在利用 LDA 模型对文本进行建模后，文本被表示成主题下的概率分布，文本向量维数降低，因而两个文本的相似度可以通过文本间的主题相似性来反应。文本的主题相似度是通过计算与文本对应的主题概率分布间的相似度来实现的[9]。文本主题概率密度之间的差距越小，则文本间相似程度越大。

一般地，最常用 KL (Kullback-Leibler Divergence)距离来衡量两个概率密度之间的差异情况，它是基于信息熵的概念定义的，也被称为相对熵、交叉熵[10]。假设 $P(x)$ 和 $Q(x)$ 是 X 上的两个概率密度函数，它们间的 KL 距离定义如下：

$$\begin{aligned} D_{KL}(P\|Q) &= \sum_{x\in X} P(x)\log\frac{1}{Q(x)} - \sum_{x\in X} P(x)\log\frac{1}{P(x)} \\ &= \sum_{x\in X} P(x)\log\frac{P(x)}{Q(x)} \end{aligned} \tag{5}$$

由公式(5)可知，KL 距离具有非负性和不对称性，即 $D_{KL}(P\|Q)\neq D_{KL}(Q\|P)$。

JS (Jensen-Shannon)距离是对 KL 距离的对称改进，并且将距离值定义在[0,1]的闭区间上[11]。因此，本文使用 JS 距离来衡量文本间的主题相似度。JS 距离定义如公式(6)所示：

$$JS(z_i\|z_j) = \frac{1}{2}D_{KL}\left(z_i\|\frac{z_i+z_j}{2}\right) + \frac{1}{2}D_{KL}\left(z_j\|\frac{z_i+z_j}{2}\right) \tag{6}$$

其中，z_i 和 z_j 分别为文本 d_i 和文本 d_j 主题概率密度。

综合以上分析，文本 d_i 和文本 d_j 间的主题相似度可表示如下：

$$sim_{JS}(d_i,d_j) = JS(z_i\|z_j) \tag{7}$$

3. 基于 LDA 模型的话题检测

3.1. 专业词典构建

在图书馆特定领域，存在一些专业词汇，对文本进行分词处理时，会将专业词汇切碎或不能识别专业词汇，这在很大程度上影响后续聚类效果。举例说明如下：

文本 d_1：图书借阅证丢失；分词结果：图书 借阅 证 丢失

文本 d_2：图书借阅期限；分词结果：图书 借阅 期限

例子中的两个短文本属于两个不同话题，但是相似度计算结果显示两者具有很大的相似性，这两个短文本在话题检测时会被归属到同一个话题。文本 d_1 中，“借阅证”一词是图书馆领域的一个专业名词，“借阅”和“证”这两个切碎的特征词在文本中是同时出现的，因此本文考虑词汇间上下文语义关系，进行上下文共现词合并，得到图书馆领域专业词汇，从而提高聚类效果。

如果两个特征词同时出现在文本集中的次数大于或等于 3，则称这两个特征词为一组专业词汇共现词对，并移入专业词典。专业词典构建方法如下所示：

算法：共现词对发现方法

输入：短文本集 $D=\{d_k\,|\,k=1,2,\cdots,N\}$

输出：专业词典 $W=\{w_{ij}\}$

Step 1：构建文本级中的词汇表 $V=\{w_j\,|\,j=1,2,\cdots,M\}$；

Step 2：遍历词汇集，去除只出现一次的词汇标记；

Step 3：将词汇表中的词汇两两合并，得到共现词对集 $W_{□}=\{w_{xy}\}$；

Step 4: for each w_{xy} from T

for each d_k from D

if $w_{xy}\subseteq d_k$

frequency(w_{xy})++

end if

end for

end for

Step 5：遍历共现词对集 $W_{总}=\{w_{xy}\}$ 中的每个共现词对，如果 frequency$(w_{xy})\le 2$，则将其去除；

Step 6：返回 frequency$(w_{xy})\ge 3$ 的共现词对。

3.2. 聚类方法

微信图书馆中的数据是按照时间顺序不断到达的动态信息流，针对这种流式数据，本文选择 Singles-Pass 聚类算法[12]用于微信文本的增量聚类。Singles-Pass 聚类算法是话题检测中最常使用的一种话题识别算法，其基本思想是依据文本的输入次序，依次处理每次输入的文本。算法具体步骤描述如下：

输入：短文本集 $D=\{d_k|k=1,2,\cdots,N\}$

输出：热点话题集 $T=\{T_i|i=1,2,\cdots,M\}$

Step 1：将第一篇到达的文本 d_1 作为初始话题类别；

Step 2：对其后新到达的文本 $d_k\ (1<k\le N)$，计算该文本与已有所有话题类别 T_i 主题相似度 $sim(d_k,T_i)=sim(d_k,d_1)+sim(d_k,d_2)+\cdots+sim(d_k,d_h)/h$，其中，$d_1,d_2,\cdots,d_h$ 为话题 T_i 内所有的文本数据；

Step 3：找出与 d_k 具有最大相似度的话题类别 T，即；$sim(d_k,T)=\arg\max sim(d_k,T_i)$；

Step 4：

a) 若 $sim(d_k,T)$ 大于给定阈值 ε，则将文本 d_k 归入对应的话题类别；

b) 若 $sim(d_k,T)$ 小于或等于给定阈值 ε，则创建一个新的话题类别，并将文本 d_k 归入到该话题类别中；

Step 5：重复步骤 2、3 和 4，直到聚类结束。

3.3. 基于 LDA 模型的热点话题检测方法

本文结合特征词上下文语义关系，构建图书馆领域专业词典，然后利用 LDA 模型对微信数据建模，应用主题相似度计算文本间的相似性，实现了微信图书馆的热点话题检测，具体步骤如下：

1) 文本预处理。合并用户在时间窗口内的文本，得到文本集合 $D=\{d_k|k=1,2,\cdots,N\}$，并对文本进行分词处理，去除停用词及标点符号。

2) 专业词典构建。根据 2.1 小节中的方法合并上下文共现词对，构建图书馆领域专业词典；

3) 主题建模。应用 LDA 模型对微信数据建模，为每一个文本都建立一个与之相对应的 LDA 模型，并使用 Gibbs 抽样算法估计每个文本的文本－主题分布参数 θ 和主题-词汇分布参数 φ。

4) 相似度计算。利用主题相似度，计算数据集中文本间的相似性；

5) 文档聚类。将当前文本和已有话题相比较，如果小于聚类阈值，则认为检测到新的话题，否则将该报道归入已有子话题类别中；

6) 得到热点话题集。对数据集下所有文档进行上述步骤，最终输出热点话题集合。

4. 实验与分析

4.1. 实验语料

本文的实验语料取自西安工业大学微信图书馆中从 2016 年 9 月到 2017 年 1 月一学期的微信文本数据。经过去除书目检索信息和合并用户在时间窗口内的文本信息后，得到 6 种类别，共 4200 篇文本信息。数据集如表 1 所示。

4.2. 评测指标

本实验主要采用文本分类中常见的准确率、召回率和 F_1 值来评估话题分类性能。

查准率： $\mathrm{Pr}=\frac{x}{z}\times 100\%$

召回率： $\mathrm{Re}=\frac{x}{y}\times 100\%$

F_1 值： $F_1=\frac{2\times \mathrm{Pr}\times \mathrm{Re}}{\mathrm{Pr}+\mathrm{Re}}\times 100\%$

其中， x 为实验数据集中正确识别的热点话题数目， y 为实验数据集中存在的热点话题数目， z 为实验数据集中识别出的热点话题数目。

4.3. 实验及结果分析

实验 1：确定 LDA 模型主题数目

利用 LDA 模型对数据集建模时，LDA 模型的性能在很大程度上受主题数 T 的影响。如果直接设定一个固定值，LDA 模型的建模效果可能会比预期值低，因此需要预先确定 LDA 模型中的最优主题数目 T。

本文利用贝叶斯统计中的标准方法[13]确定最优主题数 T，令 T 取不同的值分别运行 Gibbs 抽样算方法，查看困惑度的变化。实验结果如图 2 所示。

其中，横坐标表示主题数目，纵坐标表示 LDA 模型困惑度，图中的曲线反映了 LDA 模型困惑度随主题个数变化的趋势。可以看出，当 $T=50$ 时，LDA 模型困惑度最小，因此，本文在实验中所采用的主题数均为 50。

实验 2：构建专业词典前后系统性能对比实验

图 3 中，横坐标代表话题，纵坐标代表话题检测的准确率。本实验对文中给出的通过构建专业词典合并特征词方法进行可行性验证。两组实验均利用 LDA 模型进行文本表示，只是在建模之前，前者未作处理，后者先利用专业词典合并特征词。可以看出，通过专业词典合并特征词后，系统能够更准确的识别热点话题。

实验 3：确定聚类阈值 ε

经统计得出，实验数据集中同一话题下的文本相似度一般在 0.55 以上，而不同话题间的相似度一般

Table 1. Distribution of the hot topics under the data set
表 1. 数据集下热点话题的分布情况

编号	话题	文本数目	编号	话题	文本数目
1	开闭馆时间	718	4	图书证挂失与补办	562
2	信息绑定	1026	5	考研自习室	855
3	借阅细则	537	6	图书馆藏位置分布	502

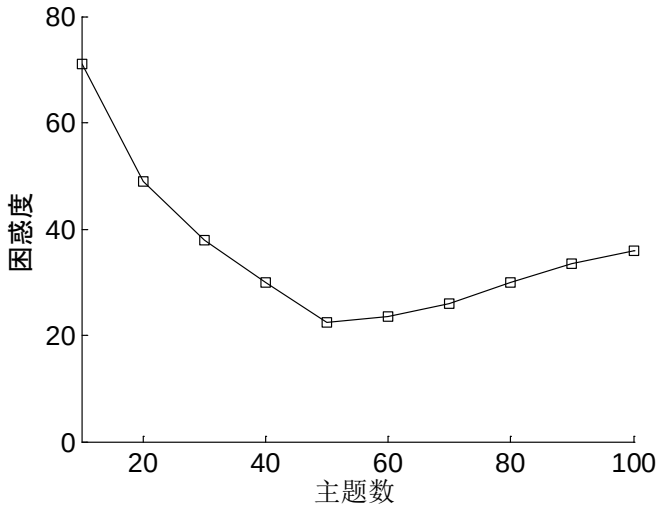

Figure 2. The trend of phenomenon of LDA model with the number of subjects
图 2. LDA 模型困惑度随主题数的变化趋势

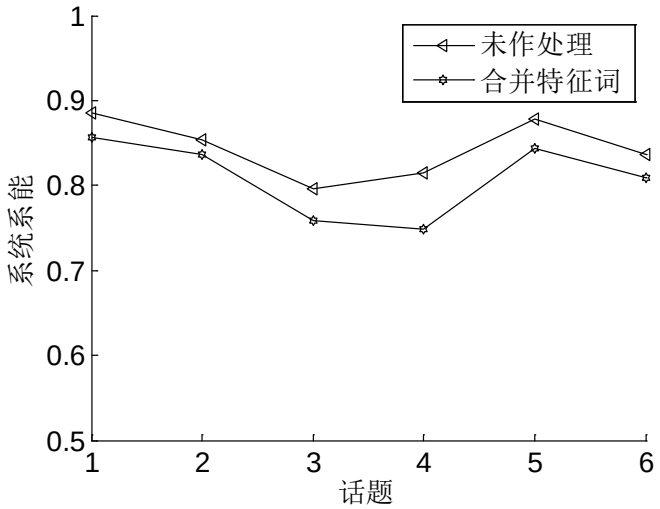

Figure 3. Comparison of the accuracy of the results of the test results
图 3. 话题检测结果准确率对比

在 0.3 以下，因此，本文设定阈值 ε 的范围为[0.32,0.54]。系统性能评价使用 6 个子话题检测代价的宏平均来衡量，结果如图 4 所示。

其中，横坐标代表预设定的聚类阈值，纵坐标代表系统的归一化开销。从图 4 可以看出，当阈值取 0.46 时，系统的归一化开销最小，因此，本文设定聚类阈值 $\varepsilon = 0.46$。

实验 4：对比试验

本文采用的 Baseline 方法为 VSM 方法，本文方法为第 2.3 节中给出的话题检测方法。实验结果如表 2 所示。

依据表 2 的统计结果，将 6 个话题类别上的查准率与召回率求平均，并对所有话题上的 F_1 值求宏平均，两种方法的对比结果如图 5 所示。

根据图 5 可以看出，在分类性能上本文方法总体上较 Baseline 方法有提高。从 F_1 值上来看，本文方法高出 3 个百分点。

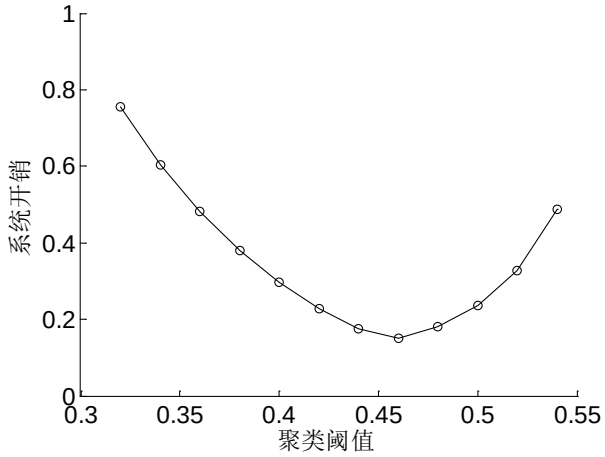

Figure 4. The trend of system normalization detection overhead changes with threshold
图 4. 系统归一化检测开销随阈值变化的趋势

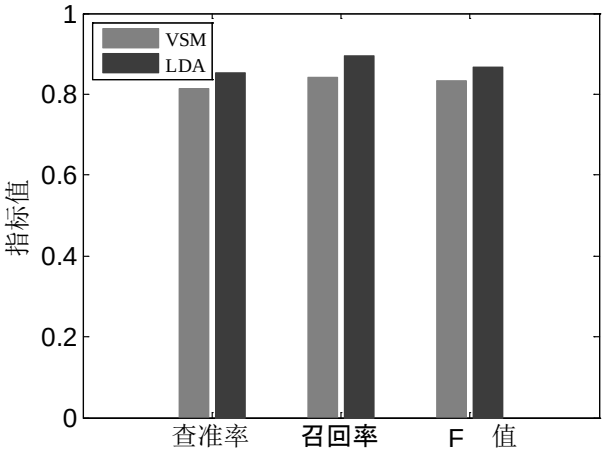

Figure 5. Comparison of the two methods
图 5. 两种方法的综合比较

Table 2. Comparison of experimental results
表 2. 实验结果对比

话题类别	VSM			LDA		
	Pr	Re	F_1 值(%)	Pr	Re	F_1 值(%)
1	0.821	0.874	84.7	0.851	0.902	87.5
2	0.806	0.828	81.6	0.837	0.893	86.4
3	0.795	0.827	81.1	0.802	0.878	83.8
4	0.816	0.853	83.4	0.862	0.897	87.9
5	0.862	0.897	87.9	0.885	0.924	90.4
6	0.784	0.846	81.3	0.812	0.869	83.9

5. 总结

本文给出的方法能够对用户发布的信息进行分类处理，获取用户所关注的话题。在此基础上，图书

馆工作人员可以根据话题检测结果做出相应的反馈，及时对用户所关注的问题发布通知公告、问题答疑或设置相应的关键字自动回复模块，不但可以节省大量的人力和物力，还能够进一步提高图书馆的服务质量。

参考文献 (References)

[1] Quan, X., Liu, G., Lu, Z., Ni, X. and Liu, W. (2010) Short Text Similarity Based on Probabilistic Topics. *Knowledge and Information Systems*, **25**, 473-491. https://doi.org/10.1007/s10115-009-0250-y

[2] 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.

[3] 孙励. 基于微博的热点话题发现[D]: [硕士学位论文]. 北京: 北京邮电大学, 2012.

[4] 刘红兵, 李文坤, 张仰森. 基于 LDA 模型和多层聚类的微博话题检测[J]. 计算机技术与发展, 2016, 26(6): 25-30.

[5] 汪进祥. 基于主题模型的微博话题挖掘[D]: [硕士学位论文]. 北京: 北京邮电大学, 2015.

[6] 余传明, 张小青, 陈雷. 基于 LDA 模型的评论热点挖掘: 原理与实现[J]. 情报理论与实践, 2010, 33(5): 103-106.

[7] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993-1022.

[8] Griffiths, T. (2002) Gibbs Sampling in the Generative Model of Latent Dirichlet Allocation.

[9] 孙昌年, 郑诚, 夏青松. 基于 LDA 的中文文本相似度计算[J]. 计算机技术与发展, 2013(1): 217-220.

[10] 方正. 微博短文本分析技术研究及应用[D]: [硕士学位论文]. 成都: 电子科技大学, 2014.

[11] 王鹏, 高铖, 陈晓美. 基于 LDA 模型的文本聚类研究[J]. 情报科学, 2015(1): 63-68.

[12] 朱恒民, 朱卫未. 基于 Single-Pass 的网络话题在线聚类方法研究[J]. 现代图书情报技术, 2011(12): 52-57.

[13] Steyvers, M. and Griffiths, T. (2007) Probabilistic Topic Models. *Handbook of Latent Semantic Analysis*, **427**, 424-440.