Journal of Computer Applications
计算机应用,2013,33(6):1587-1590
ISSN 1001-9081
CODEN JYIIDU
2013-06-01
http://www.joca.cn

文章编号:1001-9081(2013)06-01587-04
doi:10.3724/SP.J.1087.2013.01587

基于 LDA 主题模型的短文本分类方法

张志飞*,苗夺谦,高 灿
(1.同济大学 计算机科学与技术系,上海 201804; 2.同济大学 嵌入式系统与服务计算教育部重点实验室,上海 201804)
(*通信作者电子邮箱 tjzhifei@163.com)

摘 要:针对短文本的特征稀疏性和上下文依赖性两个问题,提出一种基于隐含狄列克雷分配模型的短文本分类方法。利用模型生成的主题,一方面区分相同词的上下文,降低权重;另一方面关联不同词以减少稀疏性,增加权重。采用 K 近邻方法对自动抓取的网易页面标题数据进行分类,实验表明新方法在分类性能上比传统的向量空间模型和基于主题的相似性度量分别高5%和2.5%左右。

关键词:短文本;分类;K 近邻;相似度;隐含狄列克雷分配
中图分类号:TP18 **文献标志码**:A

Short text classification using latent Dirichlet allocation

ZHANG Zhifei*, MIAO Duoqian, GAO Can
(1. *Department of Computer Science and Technology, Tongji University, Shanghai* 201804, *China*;
2. *Key Laboratory of Embedded System and Service Computing, Ministry of Education* (*Tongji University*), *Shanghai* 201804, *China*)

Abstract: In order to solve the two key problems of the short text classification, very sparse features and strong context dependency, a new method based on latent Dirichlet allocation was proposed. The generated topics not only discriminate contexts of common words and decrease their weights, but also reduce sparsity by connecting distinguishing words and increase their weights. In addition, a short text dataset was constructed by crawling titles of Netease pages. Experiments were done by classifying these short titles using K-nearest neighbors. The proposed method outperforms vector space model and topic-based similarity.

Key words: short text; classification; K-Nearest Neighbor (K-NN); similarity measure; latent Dirichlet allocation

0 引言

互联网上常见的文本数据有新闻、博客和邮件等,从这些数据中自动抽取有价值的信息和知识的技术,称为文本挖掘。文本根据长度的不同可以分为长文本和短文本两类,起初的研究并没有明确区分。随着社交媒体的兴起,移动短信、Tweet 和微博等短文本层出不穷。由于参与者多以及发布频率快,短文本的规模飞速增长。此外,短文本在搜索引擎[1]、自动问答[2]和话题跟踪[3]等领域发挥着重要的作用。短文本挖掘日益受到研究者的关注。

短文本挖掘区别于长文本挖掘,主要困难在于短文本的关键特征非常稀疏和上下文依赖性强[4]。传统的文本表示模型以及机器学习方法直接应用到短文本上效果不佳。文本表示最常用的模型是向量空间模型(Vector Space Model, VSM)[5],由于特征极度稀疏,根据词语的共现程度来衡量文本之间的相似性不再有效,导致分类或者聚类不能取得理想的效果[6-7]。目前通常的解决方法有两类:第一类是借助外部文本如搜索引擎结果,扩展短文本[8-9];第二类是借助知识库如 WordNet 或 Wikipedia 等,挖掘短文本中词语之间的内在联系[10-11]。

第一类方法不是很理想,一方面消耗较长时间,另一方面对搜索引擎的结果非常依赖;第二类方法利用知识库可以发现大部分词之间的语义关系,但是对于知识库中不存在的词无能为力。“主题”这个概念后来被引入到短文本相似性度量中,这些工作的基础是概率主题模型,以隐含狄列克雷分配(Latent Dirichlet Allocation, LDA)[12]为代表,通过隐含主题将文本关联起来。Phan 等[6]对一个大语料库使用 LDA 模型得到一簇隐含主题,然后将测试文本全部转化为主题表示形式,由于处理过程都是针对完整的数据集,该方法不适用于计算两篇文本之间的相似性。Quan 等[13]则借助主题作为第三方,考察词语之间的关联性,并进一步度量两篇文本之间的相似性。Chen 等[14]考虑主题数的影响,将单层次的主题扩展至多粒度主题。本文延续文献[13]的工作,解决其不能处理上下文依赖性的问题,实验结果表明提出的新方法取得了相对较好的结果。

1 相关工作

1.1 向量空间模型

VSM 由 Salton 等[5]提出,已经成为信息检索领域常用的文本表示模型,将文本看作“词袋”。给出一些符号定义:词表 $V=\{v_1,v_2,\cdots,v_N\}$,N 为词的总数;文本集 $D=\{d_1,d_2,\cdots,d_M\}$,M 为文本的总数;一篇文本 $d_i\in D$ 的向量表示为 $\boldsymbol{V}^{(i)}=(w_1^{(i)},w_2^{(i)},\cdots,w_N^{(i)})$,$w_k^{(i)}$ 为词 $v_k\in V$ 在 d_i 中的权重,通常采用 TF-IDF[15]权重评价函数:

收稿日期:2012-12-17;**修回日期**:2013-01-24。
基金项目:国家自然科学基金资助项目(60970061,61075056,61103067);中央高校基本科研业务费专项资金资助项目。
作者简介:张志飞(1986-),男,江苏如东人,博士研究生,CCF 学生会员,主要研究方向:粒计算、文本挖掘; 苗夺谦(1964-),男,山西祁县人,教授,博士生导师,CCF 高级会员,主要研究方向:粗糙集、Web 智能、机器学习; 高灿(1983-),男,湖南南县人,博士研究生,CCF 学生会员,主要研究方向:粗糙集、机器学习。

$$w_k^{(i)} = tf_{ki} \times \ln \frac{M}{df_k} \tag{1}$$

其中：tf_{ki} 表示 v_k 在 d_i 中出现的次数，df_k 表示 D 中含有 v_k 的文本的总数。通常采用余弦距离计算两篇文本之间的相似度[5]：

$$sim(d_i,d_j) = \frac{\sum_{k=1}^{N}(w_k^{(i)} \times w_k^{(j)})}{\sqrt{\sum_{k=1}^{N}(w_k^{(i)})^2 \times \sum_{k=1}^{N}(w_k^{(j)})^2}} \tag{2}$$

1.2 隐含狄列克雷分配

LDA 主题模型由 Blei 等[12]提出，是一个“文本—主题—词”的三层贝叶斯产生式模型，每篇文本表示为主题的混合分布，而每个主题则是词上的概率分布。最初的模型只对文本—主题概率分布引入一个超参数使其服从 Dirichlet 分布，随后 Griffiths 等[16]对主题—词概率分布也引入一个超参数使其服从 Dirichlet 分布。该模型用图 1 表示，各符号的含义如表 1 所示。

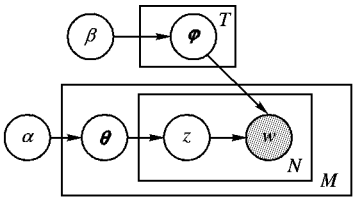

图 1 LDA 的图表示[12]

表 1 LDA 模型中各符号的含义

符号	含义	符号	含义
α	θ 的超参数	w	词
β	φ 的超参数	M	文本数
θ	文本—主题概率分布	N	词数
φ	主题—词概率分布	T	主题数
z	词的主题分配		

两个超参数一般设置为 $\alpha = 50/T, \beta = 0.01$[12]。LDA 模型的参数个数只与主题数和词数有关，参数估计是计算出文本—主题概率分布以及主题—词概率分布，即 θ 和 φ。通过对变量 z 进行 Gibbs 采样[16]间接估算 θ 和 φ：

$$\theta_{ms} = \frac{n_m^{(s)} + \alpha}{\sum_{j=1}^{T} n_m^{(j)} + T\alpha} \tag{3}$$

$$\varphi_{sk} = \frac{n_s^{(k)} + \beta}{\sum_{i=1}^{N} n_s^{(i)} + N\beta} \tag{4}$$

其中：$n_m^{(j)}$ 表示文本 d_m 中赋予主题 j 的词的总数，$n_s^{(i)}$ 表示词 v_i 被赋予主题 s 的总次数。

1.3 基于主题的相似性

Quan 等[13]提出了基于主题的相似性（Topic-Based Similarity，TBS）度量方法来解决短文本的特征稀疏性问题，基本思想是通过第三方主题来比较两篇短文本。

假设文本集 D 中存在两篇短文本 d_1 和 d_2，它们的向量表示为 $\boldsymbol{V}^{(1)} = (w_1^{(1)}, w_2^{(1)}, \cdots, w_N^{(1)})$ 和 $\boldsymbol{V}^{(2)} = (w_1^{(2)}, w_2^{(2)}, \cdots, w_N^{(2)})$。在 D 上运行 LDA 模型后得到 T 个隐含主题以及主题—词概率分布 φ，记 φ_{sk} 为词 v_k 属于主题 $s(1 \leqslant s \leqslant T)$ 的概率。

这两篇短文本的可区分词集[13]定义为：

$$\begin{cases} Dist(d_1) = \{v \mid v \in d_1 \wedge v \notin d_2\} \\ Dist(d_2) = \{v \mid v \in d_2 \wedge v \notin d_1\} \end{cases} \tag{5}$$

对于所有主题 s，寻找可区分词集中主题—词概率值最大的词 $v_m \in Dist(d_1)$ 和 $v_n \in Dist(d_2)$，其值为 φ_{sm} 和 φ_{sn}。如果它们均不小于阈值 λ，则认为 v_m 和 v_n 在主题 s 上非常相关，增加在各自文本中的权重[13]：

$$\begin{cases} w_n^{(1)} = w_n^{(1)} + w_n^{(2)} \times \varphi_{sn} \\ w_m^{(2)} = w_m^{(2)} + w_m^{(1)} \times \varphi_{sm} \end{cases} \tag{6}$$

2 基于 LDA 的短文本分类

2.1 问题描述

针对特征稀疏性问题，借助隐含主题关联不同的词，减少稀疏性对相似性度量的影响。给一个例子，如图 2 所示。

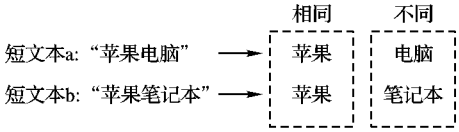

图 2 短文本例子（特征稀疏性）

图 2 中“电脑”和“笔记本”是不同的词，TBS 引入隐含主题，将两者很强地关联起来，认为两者是同义词。

另外一个例子体现了短文本上下文依赖性强的问题，如图 3 所示。

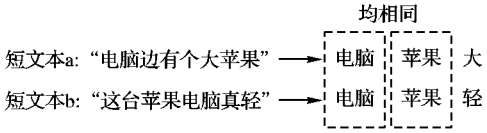

图 3 短文本例子（上下文依赖性）

对这个例子中的短文本分词并去除停用词后得到右边所示的词集。VSM 会认为这两篇文本具有很大的相似性；TBS 对可区分词“大”和“轻”的处理对相似性度量影响不大。因此，TBS 不能处理上下文依赖性问题，或者说是多义词问题。

基于以上考虑，提出了一种新的基于 LDA 的短文本分类方法，延续了 TBS 中第三方主题的思想，并进一步利用文本—主题概率分布来解决上下文依赖性问题。

2.2 相似性度量

TBS 利用 LDA 模型的主题—词概率分布解决特征稀疏性问题，新方法则再利用 LDA 模型的文本—主题概率分布解决上下文依赖性问题。

除了可区分词集外，还定义共有词集：

$$Comm(d_1,d_2) = \{v \mid v \in d_1 \wedge v \in d_2\} \tag{7}$$

对共有词集中同时满足条件 C1 和 C2 的词降低权重来解决上下文依赖性问题。前者通过最大主题判断两篇文本是否属于同一主题，属于不同主题的相同词很有可能是多义词；后者通过概率值排名忽略和主题关系不明显的词。

条件 C1：根据式(8)，提取 d_1 和 d_2 各自的最大主题，两者不一致；

条件 C2：在各自的最大主题下，该词的主题—词概率值排名前 60%。

$$\begin{cases} t_{\max}^{(1)} = \arg\max_s \{\theta_{1s} \mid 1 \leqslant s \leqslant T\} \\ t_{\max}^{(2)} = \arg\max_s \{\theta_{2s} \mid 1 \leqslant s \leqslant T\} \end{cases} \tag{8}$$

类比式(6)，采用式(9)降低满足条件的共有词 v_c 的权重：

$$\begin{cases} w_c^{(1)} = \left| w_c^{(1)} - w_c^{(2)} \times \theta_{1t_{\max}^{(1)}} \times \varphi_{t_{\max}^{(1)}c} \right| \\ w_c^{(2)} = \left| w_c^{(2)} - w_c^{(1)} \times \theta_{2t_{\max}^{(2)}} \times \varphi_{t_{\max}^{(2)}c} \right| \end{cases} \tag{9}$$

基于 LDA 的短文本分类方法中的相似性度量新算法描述如下:

算法　相似性度量新算法。

输入　短文本 d_1 和 d_2,概率分布 $\boldsymbol{\varphi}$ 和 $\boldsymbol{\theta}$;

输出　相似度 $sim(d_1,d_2)$。

第 1 步　获取共有词集 $Comm(d_1,d_2)$;

第 2 步　根据式(8)提取两篇文本各自的最大主题 $t_{\max}^{(1)}$ 和 $t_{\max}^{(2)}$,如果 $Comm(d_1,d_2) = \varnothing$ 或者 $t_{\max}^{(1)} = t_{\max}^{(2)}$,则转至第 4 步;

第 3 步　对于共有词集中的每个词 v_c,如果还满足条件 C2,则根据式(9)更新权重;

第 4 步　获取可区分词集 $Dist(d_1)$ 和 $Dist(d_2)$,如果有一个为空,则转至第 6 步;

第 5 步　对于每个主题 s,分别找出 $Dist(d_1)$ 和 $Dist(d_2)$ 中具有最大主题—词概率值的词 v_m 和 v_n,其概率值为 φ_{sm} 和 φ_{sn},如果 $\varphi_{sm} \geqslant \lambda$ 且 $\varphi_{sn} \geqslant \lambda$,则根据式(6)更新权重;

第 6 步　根据式(2)计算 d_1 和 d_2 的相似度 $sim(d_1,d_2)$。

该算法和 TBS 相比,增加的时间开销在于第 1 步至第 3 步,其中提取最大主题的时间复杂度是 $O(T)$,共有词集最大长度为 N,权重更新解决上下文依赖性问题的时间复杂度为 $O(N)$,因此新算法增加的时间复杂度为 $O(T+N)$。

2.3　分类框架

分类的整体框架如图 4 所示,A 部分在训练文本集上运行 LDA 模型,得到隐含主题以及主题—词概率分布;B 部分针对新文本在已生成的隐含主题上运行 LDA 模型,得到文本—主题概率分布;C 部分则结合主题—词概率分布以及各自的文本—主题概率分布计算两篇文本之间的相似性,并送入分类器。

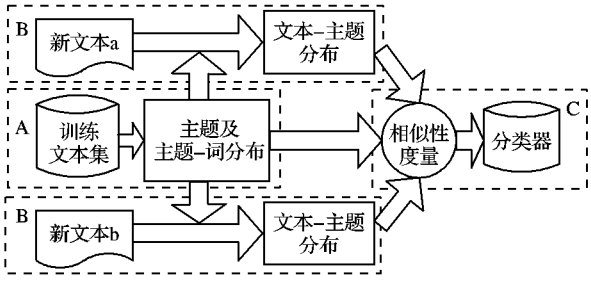

图 4　分类框架

3　实验

3.1　实验数据

生活中有很多的短文本数据,如短信、邮件、评论和新闻标题等。短信、邮件和评论等一般属于二值分类问题,此处使用新闻标题作为短文本数据,在多类别分类问题下开展实验。

采用爬虫自动抓取网易网页,页面包括标题和正文,将页面标题作为短文本,所属的分类作为文本类别。数据集如表 2 所示,文本共有 8 种类别,共计 5 892 篇。

3.2　实验设置

预处理　对抓取的原始文本进行预处理,包括 HTML 解析、分词和去除停用词,其中分词采用了 ICTCLAS 系统。

主题数　将网易页面正文数据的 2/3 用于学习,1/3 用于预测。利用困惑度 Perplexity 指标[12]确定主题数。该指标表示预测数据时的不确定度,取值越小表示性能越好。

表 2　数据集

类别	文本数	类别	文本数
教育	518	商务	501
经济	701	社会	483
军事	1 871	体育	808
科技	505	娱乐	505

TBS 阈值　TBS 调整可区分词的权重时需要判断主题—词概率值是否超过阈值 λ。利用最大主题—词概率值自动确定阈值,将所有主题下的最大主题—词概率值累加求平均,并以 60% 的分界作为 λ 的取值:

$$\lambda = 0.6 \times \frac{1}{T} \sum_{i=1}^{T} \max_{1 \leqslant j \leqslant N} \varphi_{ij} \tag{10}$$

分类器　采用 K 近邻(K-Nearest Neighbors,K-NN)方法对页面标题分类并进行五折交叉验证。K-NN 中的近邻数为 VSM 取得最好性能时的近邻数,在同一近邻数下,比较不同方法的分类效果。

评估　采用文本分类中常见的指标来评估分类性能,查全率 Re 和查准率 Pr 以及两者的综合评价 F_1 值[5]:

$$F_1 = \frac{2 \times Pr \times Re}{Pr + Re} \tag{11}$$

3.3　实验结果

主题数　将 LDA 模型的主题数设置为 10 ~ 100(间隔 10),三次实验的困惑度随主题数变化情况如图 5 所示。主题数不断增加,困惑度呈逐渐下降趋势,当达到 50 时,下降趋势不再明显。主题数越多,LDA 模型估计的参数越多,计算代价越大,因此取 $T = 50$。

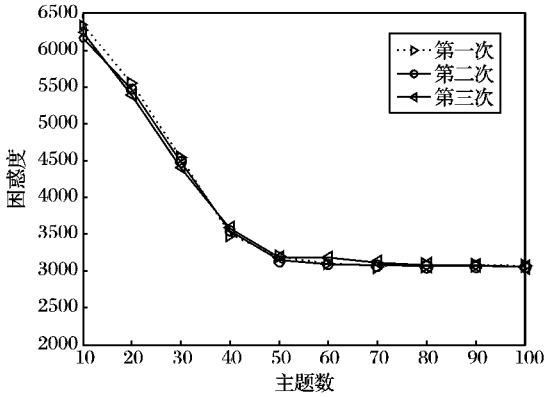

图 5　不同主题数下的困惑度

近邻数　采用 VSM 表示标题数据集,将 K-NN 中的近邻数设置为 11 ~ 25(间隔 2)。不同近邻数下的 8 个类别的 F_1 值如图 6 所示,将近邻数设为 17 最为合理。

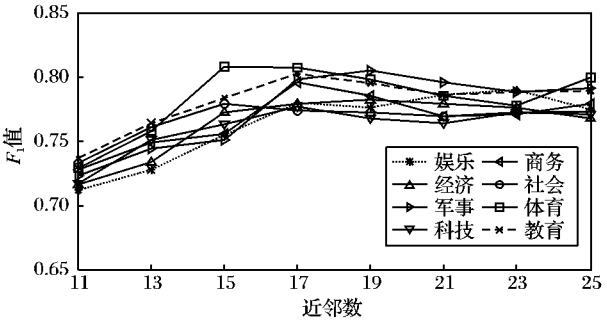

图 6　不同近邻数下的分类性能

查全率和查准率比较:设置主题数为 50,K-NN 的近邻数为 17,三种方法在 8 个类别上的查全率和查准率如表 3 所示。

新方法在各类别上的查全率和查准率均最优,TBS次之,VSM最差。

表3 三种方法的查全率和查准率

类别	VSM		TBS		本文方法	
	Re	*Pr*	*Re*	*Pr*	*Re*	*Pr*
教育	0.782	0.779	0.801	0.802	0.824	0.831
经济	0.778	0.780	0.792	0.791	0.812	0.815
军事	0.803	0.794	0.820	0.823	0.829	0.830
科技	0.776	0.778	0.795	0.798	0.821	0.820
商务	0.791	0.802	0.812	0.816	0.820	0.821
社会	0.772	0.776	0.799	0.801	0.825	0.827
体育	0.812	0.803	0.820	0.823	0.838	0.840
娱乐	0.804	0.803	0.821	0.822	0.837	0.839

同/多义词分布:为了解释方法优劣的原因,统计同义词和多义词的出现情况。同义词对应可区分词集中更新过权重的词,而多义词对应共有词集中更新过权重的词。同义词和多义词个数如表4所示。

表4 同/多义词分布

类别	同义词个数	多义词个数
教育	251	85
经济	144	126
军事	182	26
科技	236	132
商务	178	31
社会	240	109
体育	119	122
娱乐	239	94

由表4可知,短文本确实存在上下文依赖性的问题。虽然多义词规模不大,但会一定程度上影响短文本的相似性度量。另外,军事类和商务类的多义词个数较少,这也正好解释了表3中新方法在这两类上优势不明显的原因。

综合比较:将8个类别上的查全率和查准率求平均,进而得到F_1(准确描述应该为宏F_1),三种方法的对比结果如图7所示。

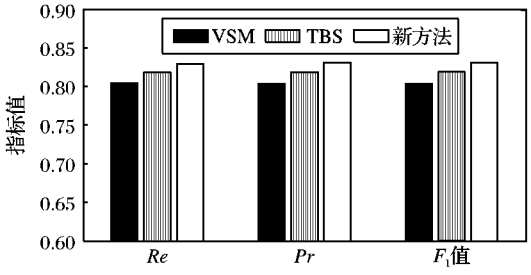

图7 三种方法的综合比较

显然,在分类性能上新方法优于TBS,也优于VSM。从F_1值上来看,新方法超出TBS约2.5个百分点,超出VSM约5个百分点。新方法在不增加时间代价的前提下取得更好的分类精度。

4 结语

短文本处理时面临两个问题:特征稀疏性和上下文依赖性。利用LDA模型生成主题,TBS方法只是解决了特征稀疏性问题。因此,提出新方法进一步解决了上下文依赖性问题。新方法不仅给出短文本相似性的完备度量,而且能够自动确定TBS阈值。对网易页面标题数据进行*K*-NN分类,通过实验设置合理的主题数和近邻数,结果表明新方法的分类性能优于TBS和VSM,并通过同义词和多义词分布解释了其原因。社交媒体由于表述的口语化和不规范化,给短文本处理带来新的挑战。

参考文献:

[1] PARK E K, RA D Y, JANG M G. Techniques for improving Web retrieval effectiveness [J]. Information Processing Management, 2005, 41(5): 1207 - 1223.

[2] LIU W Y, HAO T Y, CHEN W, *et al*. A Web-based platform for user-interactive question-answering[J]. World Wide Web, 2009, 12(2): 107 - 124.

[3] 郑斐然, 苗夺谦, 张志飞, 等. 一种中文微博新闻话题检测的方法[J]. 计算机科学, 2012, 39(1): 138 - 141.

[4] 贺涛, 曹先彬, 谭辉. 基于免疫的中文网络短文本聚类算法[J]. 自动化学报, 2009, 35(7): 896 - 902.

[5] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing [J]. Communications of the ACM, 1975, 18(11): 613 - 620.

[6] PHAN X H, NGUYEN M L, HORIGUCHI S. Learning to classify short and sparse text & Web with hidden topics from large-scale data collections[C]// Proceedings of the 17th Conference on World Wide Web. New York: ACM, 2008: 91 - 100.

[7] WANG L, JIA Y, HAN W H. Instant message clustering based on extended vector space model[C]// Proceedings of the 2nd International Conference on Advances in Computation and Intelligence. Berlin: Springer-Verlag, 2007: 435 - 443.

[8] SAHAMI M, HEILMAN T D. A Web - based kernel function for measuring the similarity of short text snippets[C]// Proceedings of the 15th Conference on World Wide Web. New York: ACM, 2006: 377 - 386.

[9] YIH W, MEEK C. Improving similarity measures for short segments of text[C]// Proceedings of the 22nd Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2007: 1489 - 1494.

[10] 翟延冬, 王康平, 张东娜. 一种基于WordNet的短文本语义相似性算法[J]. 电子学报, 2012, 40(3): 617 - 620.

[11] BANERJEE S, RAMANATHAN K, GUPTA A. Clustering short texts using Wikipedia[C]// Proceedings of the 30th Annual International ACM SIGIR Conference on on Research and Development in Information Retrieval. New York: ACM, 2007: 787 - 788.

[12] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(3): 993 - 1022.

[13] QUAN X J, LIU G, LU Z, *et al*. Short text similarity based on probabilistic topics[J]. Knowledge Information System, 2010, 25(3): 473 - 491.

[14] CHEN M, JIN X, SHEN D. Short text classification improved by learning multi-granularity topics[C]// Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2011: 1776 - 1781.

[15] SALTON G, YANG C S. On the specification of term values in automatic indexing [J]. Journal of Documentation, 1973, 29(4): 351 - 372.

[16] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(S1): 5228 - 5235.