Journal of Computer Applications
计算机应用,2014,34(11):3144-3146,3151
ISSN 1001-9081
CODEN JYIIDU
2014-11-10
http://www.joca.cn

文章编号:1001-9081(2014)11-3144-03
doi:10.11772/j.issn.1001-9081.2014.11.3144

# 基于主题的 Web 文本聚类方法

张万山,肖　瑶,梁俊杰*,余敦辉
(湖北大学 计算机与信息工程学院,武汉 430062)
(*通信作者电子邮箱 416356541@qq.com)

**摘　要**:针对传统 Web 文本聚类算法没有考虑 Web 文本主题信息导致对多主题 Web 文本聚类结果准确率不高的问题,提出基于主题的 Web 文本聚类方法。该方法通过主题提取、特征抽取、文本聚类三个步骤实现对多主题 Web 文本的聚类。相对于传统的 Web 文本聚类算法,所提方法充分考虑了 Web 文本的主题信息。实验结果表明,对多主题 Web 文本聚类,所提方法的准确率比基于 *K*-means 的文本聚类方法和基于《知网》的文本聚类方法要好。

**关键词**:多主题;Web 文本;聚类;特征词;准确率
**中图分类号**:TP391.1　　**文献标志码**:A

## Web text clustering method based on topic

ZHANG Wanshan, XIAO Yao, LIANG Junjie*, YU Dunhui
(*School of Computer and Information Engineering, Hubei University, Wuhan Hubei* 430062, *China*)

**Abstract:** Concerning that the traditional Web text clustering algorithm without considering the Web text topic information leads to a low accuracy rate of multi-topic Web text clustering, a new algorithm was proposed for Web text clustering based on the topic theme. In the method, multi-topic Web text was clustered by three steps: topic extraction, feature extraction and text clustering. Compared to the traditional Web text clustering algorithm, the proposed method fully considered the Web text topic information. The experimental results show that the accuracy rate of the proposed algorithm for multi-topic Web text clustering is higher than the text clustering method based on *K*-means or HowNet.

**Key words:** multi-topic; Web text; clustering; characteristic word; accuracy

## 0　引言

随着 Web 技术迅猛发展,互联网信息呈爆炸式增长,提供一种有效机制组织管理大规模文档,帮助用户获取有用信息变得日益急切。文本聚类作为一种重要的文本分析方法,已得到广泛研究[1-3]。

传统的文本聚类算法包括 *K*-means 及其改进算法[4-6],基于《知网》(HowNet)的文本聚类算法[7-8]等。但这些算法均没有考虑文本的主题信息,现有的考虑主题信息的 Web 文本聚类方法也相对较少,多主题 Web 文本聚类技术不够成熟,需要进一步研究。文献[9]提出一种结合语言学特征的索引聚类(Linguistic Features Indexing Clustering, LFIC)方法,该方法通过"主题元素"构建索引,从而实现文本聚类,但对如何获取"主题元素",文章没有给出相应的方法。文献[10]提出一种基于主题的 Web 文本聚类算法——HTBC,该算法提取 Web 文本中包含特征词最多的主题,并以此主题表征 Web 文本,在此基础上对 Web 文本按主题聚类,该算法对单一主题的 Web 文本聚类,能取得较好的聚类效果,但是对多主题 Web 文本聚类,聚类结果的准确率降低。文献[11]针对此提出一种基于潜在狄利克雷分布(Latent Dirichlet Allocation,LDA)模型的文本聚类研究方法,该方法首先获取整个文本集隐含的所有主题,然后将文本表示为这些主题的特定比例的混合,最后再将文本聚类,因此,该方法性能受文本集自身的影响很大,稳定性有待提高。

为此,本文提出一种基于主题的 Web 文本聚类(Theme-based Web Text Clustering,TWTC)方法。该方法利用《知网》外部语义知识库获取 Web 文本的主题,而不依赖于文本集合,因而具有较好的稳定性,同时该算法基于多主题对 Web 文本聚类,保证了聚类结果的准确性。

## 1　TWTC 算法实现流程

Web 文本聚类是 Web 挖掘的一项重要研究内容,通过 Web 文本聚类可以将大量的文本信息进行自动分类,根据文本语义划分到不同类别,以方便文本集的使用和管理。Web 文本聚类的一般过程如图 1 所示。

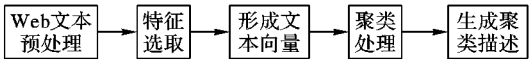


图 1　Web 文本聚类一般过程

由图 1 可知,Web 文本聚类算法首先基于 Web 预处理,选取 Web 文本的特征词,通常选取出现频率较高的词作为特征词;然后,利用选取的特征词形成 Web 文本向量;最后对 Web 文本向量进行聚类处理,并生成聚类描述。

**收稿日期**:2014-06-05;**修回日期**:2014-08-18。　**基金项目**:国家自然科学基金资助项目(61272111, 61202031, 61273216, 61202032);湖北省自然科学基金资助项目(2013CFB002, 2013CFA115);武汉市科技攻关计划项目(201210621214, 201210421132)。

**作者简介**:张万山(1973-),男,湖北武汉人,硕士,主要研究方向:Web 信息挖掘;　肖瑶(1987-),女,湖北武汉人,硕士,主要研究方向:Web 信息挖掘;　梁俊杰(1974-),女,湖北武汉人,副教授,博士,主要研究方向:数据分析、云计算;　余敦辉(1974-),男,湖北武汉人,副教授,博士,主要研究方向:服务计算、大数据。

这一过程没有充分考虑文本的主题信息,对于包含多个主题的文本的聚类结果不甚理想。

为此,本文提出 TWTC 算法,其聚类过程如图 2 所示,首先对 Web 文本进行预处理,然后抽取 Web 文本的主题,再针对主题选取特征词并形成 Web 文本主题特征向量,最后进行聚类处理和生成聚类描述。

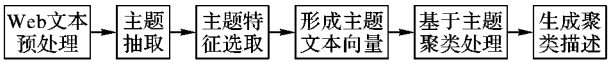


图 2　TWTC 聚类过程

## 2　TWTC 文本聚类算法具体实现

针对选中的 Web 文本,必须进行预处理,其过程是:首先对 Web 页面去噪[12],过滤掉无关信息,使 Web 文本转化为纯文本。然后利用已有的分词工具,如中国科学院的汉语词法分析系统——ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System),对转化后的 Web 纯文本进行分词,去掉其中的停用词等虚词,仅保留名词,并统计词频,从而得到"主题词集 $I$"。图 3 所示为提取"主题词集 $I$" 的过程。

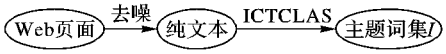


图 3　提取主题词集

TWTC 算法通过对"主题词集 $I$" 的处理,实现对 Web 文本的聚类。TWTC 算法具体分为基于《知网》提取主题、基于词频-逆文本词频(Term Frequency-Inverse Document Frequency, TF-IDF)抽取特征词集 $M$、文本聚类 3 个步骤。

### 2.1　基于《知网》的主题提取

基于《知网》对主题词集 $I$ 聚类,得到 $c$ 个簇,从中选取权重最大的 $k$ 个簇 $T_1,T_2,\cdots,T_k$ 作为文本 $D$ 的主题集合 $T$,具体实现流程见图 4。

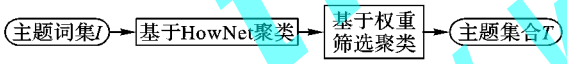


图 4　基于《知网》提取主题

具体算法描述如下:

输入　主题词集 $I=\{(w_1,f_1),(w_2,f_2),\cdots,(w_n,f_n)\}$,其中,$w_i$ 是主题词,$f_i$ 是主题词 $w_i$ 的词频。

输出　文本 $D$ 的主题 $T=\{T_1,T_2,\cdots,T_k\}$。

具体步骤:

1) 抽取主题词集 $I$ 中的非低频词,得 $I_h=\{w_1,w_2,\cdots,w_Q\}$,抽取 $I$ 中的低频词(低于词频阈值 $\theta$ 的词),得 $I_l=\{w_{Q+1},\cdots,w_n\}$。

2) 基于《知网》计算 $I_h$ 中主题词的两两相似度,得到一个 $Q\times Q$ 的相似度矩阵:

$$S=\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1Q} \\ s_{21} & s_{22} & \cdots & s_{2Q} \\ \vdots & \vdots & & \vdots \\ s_{Q1} & s_{Q2} & \cdots & s_{QQ} \end{bmatrix}$$

其中:$s_{ij}=sim(w_i,w_j);i,j\in[1,Q]$。

3) 基于相似度矩阵,对 $I_h=\{w_1,w_2,\cdots,w_Q\}$ 聚类:

a) 任选一个非低频词 $w_i\in I_h$,对 $w_j\in I_h-\{w_i\}$,查找相似度矩阵,若 $s_{ij}\geqslant\alpha$($\alpha$ 为阈值),则将 $w_i$、$w_j$ 聚类。从 $I_h$ 中删除已聚类的词。

b) 重复步骤 a),直到 $I_h$ 为空。

4) 步骤 3) 聚类可得到 $c$ 个簇,再将低频词 $I_l=\{w_{Q+1},\cdots,w_n\}$ 聚类到这 $c$ 个簇中:

a) 任选一个低频词 $w_i\in I_l$,计算 $w_i$ 与 $c$ 个簇的相似度,并将 $w_i$ 聚类到相似度最大的簇。从 $I_l$ 中删除 $w_i$。

b) 重复步骤 a),直到 $I_l$ 为空。

5) 计算各个簇 $T_r$ 的权重,记为:

$$W(T_r)=\sum_{T_r}f_i\Big/\sum_{I}f_i;\ r\in[1,c]$$

其中:$\sum_{T_r}f_i$ 是簇 $T_r$ 中所有词频之和,$\sum_{I}f_i$ 是文本 $D$ 的主题词集 $I$ 中所有词频之和。

6) 取权重最大的 $k$ 个簇,得到文本 $D$ 的主题集合:

$$T=\{T_1,T_2,\cdots,T_k\};\ k\in[1,c]$$

### 2.2　基于 TF-IDF 的特征词集抽取

依据主题 $T_1,T_2,\cdots,T_k$ 的归一化权重,计算各个主题中应该抽取特征词的个数,权重越大的主题,从中抽取特征词的个数越多。依次从主题 $T_1,T_2,\cdots,T_k$ 中抽取权重最大的 $m_1,m_2,\cdots,m_k$ 个特征词,形成文本 $D$ 的特征词集 $M$,实现流程如图 5 所示。


图 5　基于 TF-IDF 的特征词集抽取

具体算法描述如下:

1) 各个主题 $T_r$ 权重 $W(T_r)$ 归一化:

$$W_n(T_r)=\frac{W(T_r)}{\sqrt{\sum_{r=1}^{k}W^2(T_r)}};\ r\in[1,k]$$

2) 计算主题 $T_r$ 中应该抽取的特征词个数 $m_r$:

$$m_r=|M|*W_n(T_r)$$

其中:$|M|$ 是特征词集 $M$ 的个数,$W_n(T_r)$ 是主题 $T_r$ 的归一化权重。

3) 计算主题 $T_r$ 中主题词 $w_i$ 的权重 $f(w_i)$:

$$f(w_i)=V(w_i)\times\lg(N/N_k+\lambda)$$

其中:$w_i$ 表示主题 $T_r$ 中的主题词;$V(w_i)$ 表示主题词 $w_i$ 的词频;$N$ 表示文本训练集的个数;$N_k$ 表示包含主题词 $w_i$ 的文本的个数;$\lambda$ 是权重的调节因子,本文中其值为 0.01。

4) 依次从主题 $T_1,T_2,\cdots,T_k$ 中抽取权重最大的 $m_1,m_2,\cdots,m_k$ 个特征词,得文本 $D$ 的特征词集:

$$M=\bigcup_{i=1}^{m_1}w_i\bigcup_{j=1}^{m_2}w_j\cdots\bigcup_{t=1}^{m_k}w_t$$

### 2.3　文本聚类

1) 文本表示模型。本文采用如下方式表示文本 $D$:

$$D=\bigcup_{w_i\in M}(w_i,W_n(T_r))$$

其中:$w_i$ 表示文本 $D$ 的特征词,$W_n(T_r)$ 表示特征词 $w_i$ 所属的主题 $T_r$ 的归一化权重。考虑到主题 $T_r$ 的权重对文本的聚类产生的影响更大,因此,本文在表示文本 $D$ 时采用了主题对应的权重,而非特征词自身的权重。

2) 计算两文本 $D_1,D_2$ 的相似度:

$$sim(D_1,D_2)=\frac{\sum sim(w_i,w_j)\times W_{1n}(T_i)\times W_{2n}(T_j)}{\sum sim(w_i,w_j)}$$

其中：$sim(w_i,w_j)$ 是文本 $D_1$、$D_2$ 中任意两个特征词 $w_i$、$w_j$ 基于《知网》的相似度，$W_{1n}(T_i)$ 是文本 $D_1$ 的特征词 $w_i$ 所对应主题 $T_i$ 的归一化权重，$W_{2n}(T_j)$ 是文本 $D_2$ 的特征词 $w_j$ 所对应的主题 $T_j$ 所对应的归一化权重。

3）对文档进行聚类。如果 $sim(D_1,D_2) \geqslant \gamma$（$\gamma$ 是相似度阈值），则将文本 $D_1$、$D_2$ 聚类，记为：

$$C = \bigcup_{w_i \in (D_1 \cup D_2)} (w_i, W_n(T_i)) \quad (1)$$

从式(1)可知：随着聚类文本数量的增大，聚类 $C$ 中的特征词个数会增多，聚类特征向量维数会增大。可以采用权重最大的 $p$ 个特征词表示该聚类，从而降低特征向量的维数。

## 3 算法性能分析

该算法性能主要受 4 个因素影响：主题词集的规模 $n$、非低频词个数 $Q$、主题个数 $k$ 和特征词个数 $|M|$。

本文的时间复杂度由 3 部分组成：

1）主题抽取，时间复杂度为 $O(Q^2/2) + O(n-Q)$；

2）特征抽取，时间复杂度为 $O(k)$；

3）文本聚类，时间复杂度为 $O(|M|^2/2)$。

主题抽取过程中，本文没有计算所有主题词的两两相似度，而是先计算非低频词的两两相似度，将非低频词聚类后，再计算低频词与聚类之间的相似度。由于非低频词的个数 $Q$ 通常远小于主题词规模 $n$，所以 1）中时间复杂度近似为 $O(n)$。在特征抽取和文本聚类两个过程中，主题个数 $k$ 和特征词个数 $|M|$ 都远小于主题词集的规模 $n$。

通过上述分析可知，本文算法复杂度为 $O(Q^2/2) + O(n-Q) + O(k) + O(|M|^2/2) \approx O(n)$，即本文的时间复杂度近似为线性时间复杂度 $O(n)$。

## 4 实验与分析

为了验证 TWTC 算法的有效性和合理性，设计实验，将 TWTC 算法与 *K*-means 算法[4]、基于《知网》的文本聚类(Text Clustering based on HowNet，TCH)算法[7]及 HTBC[10]算法进行对比，并引入准确率和召回率两个指标作为算法性能评价指标。

准确率 $Precision(i,r)$ 和召回率 $Recall(i,r)$[13] 的定义如下：

$$Precision(i,r) = n(i,r)/n_r$$

$$Recall(i,r) = n(i,r)/n_i$$

其中：$n(i,r)$ 是聚类 $r$ 中包含的类别 $i$ 中的文档个数，$n_r$ 是聚类 $r$ 中文档的个数，$n_i$ 是预定义类别 $i$ 的文档的个数。

整个算法的准确率和召回率分别定义为各个主题准确率和各个主题召回率的加权平均值：

$$Precision = \sum_{i=1}^{N_T} \frac{n_i}{N} Precision(i,r)$$

$$Recall = \sum_{i=1}^{N_T} \frac{n_i}{N} Recall(i,r)$$

其中：$N$ 是所有主题文档的总个数，$N_T$ 是聚类的总个数。

实验一　在实验中，选择信息技术、教育、军事、文化、体育、旅游、财经、医药 8 个种类共 50 个主题，从网上搜索，随机下载3 000篇相关文本作为测试语料。在每个算法中，特征向量维度均取 50 维；TCH 算法和 TWTC 算法中，特征词的相似度阈值取 0.5；TWTC 算法中，对于主题数多于 3 个的文本，提取权重最大的 Top-2 个簇作为其主题。考虑到 *K*-means 算法的不稳定性，实验结果取 5 次运行结果的平均值。在此情况下，各算法的准确率和召回率如表 1 所示。

表 1　4 种算法的准确率和召回率比较　%

| 算法 | 准确率 | 召回率 |
|---|---|---|
| *K*-means | 58 | 82 |
| TCH | 79 | 84 |
| HTBC | 85 | 78 |
| TWTC | 96 | 85 |

从表 1 可看出：TWTC 算法的准确率远高于 *K*-means 算法和基于《知网》的文本聚类算法，较 HTBC 算法的准确率也有所提高，但召回率却高于 HTBC 算法。

实验二　通过改变测试文档的数量，比较 4 种算法准确率和召回率随文档数量增大的变化情况，从而测试 4 种算法的稳定性。测试语料文档总数据量依次选取为 3 000，6 000，12 000，24 000，对 4 种算法聚类结果进行比较，结果如图 6 所示。

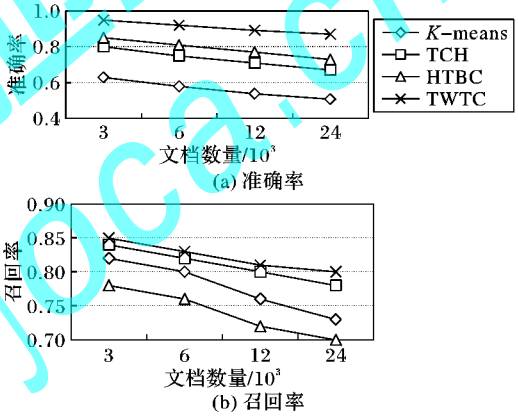


图 6　4 种算法的准确率和召回率与文档数量关系

从图 6 不难看出，随着测试文档数量的增加，四种聚类算法的准确率和召回率均有所下降，但 TWTC 算法的准确率和召回率相对于其他算法下降趋势缓慢，算法稳定性更好。

## 5 结语

本文提出一种基于主题的 Web 文本聚类方法。该方法首先基于《知网》提取 Web 文本主题；然后基于 TF-IDF 抽取特征词，并形成 Web 文本的主题特征向量；最后，借助于该向量对 Web 文本聚类。实验结果表明，采用该方法对 Web 文本进行聚类，聚类结果的准确率有一定的提高；随着测试文档数量的增加，该算法的准确率和召回率呈缓慢下降趋势，即该算法的稳定性较好。下一步的工作将对 TWTC 算法中相关阈值的确定方法进行改进和优化，从而使得该方法各项指标进一步提高。

**参考文献：**

[1] MENG X. Research on Web text clustering and retrieval technology [D]. Harbin: Harbin Institute of Technology, 2009: 1 - 10.（孟宪军. 互联网文本聚类与检索技术研究[D]. 哈尔滨：哈尔滨工业大学，2009：1 - 10.）

[2] LI Y. Text document clustering based on frequent word meaning sequences [J]. Data and Knowledge Engineering, 2008, 64(1): 381 - 404.

（下转第 3151 页）

已很少,这就说明节点基本上都参与了数据块的上传,积极地参与资源上传,达到了激励节点上传数据块的目的。

## 4　结语

本文针对 P2P 网络中节点搭便车的行为,着重研究了 P2P 流媒体点播中的节点激励机制,并提出了将歧视性第二价格拍卖应用于 P2P 流媒体点播激励机制的方案。通过仿真实验,证明本文方案在节点收益方面优于第一价格的方案,在节点上传带宽利用率上优于以最大上传带宽上传的方案,此外,系统中贡献节点的比例明显大于自私节点,自私节点数目明显减少,达到了激励节点积极参与数据块分享的效果。

本文未考虑到节点的诚信问题和方案在非实时 P2P 应用中的效果,在下一步的研究工作中可以分为两方面:一方面针对节点的诚信问题作出约束,规范拍卖市场;另外一方面对比本文方案在实时性与非实时性应用中的效果,验证方案的有效性。

**参考文献:**


[1] HUGHES D, COULSON G, WALKERDINE J. Free riding on Gnutella revisited: the bell tolls?[EB/OL]. [2013-01-22]. http://www.computer.org/csdl/mags/ds/2005/06/o6001.pdf

[2] GHEORGHIU S, LIMA L, TOLEDO A L, *et al*. A layered network coding solution for incentives in peer-to-peer live streaming[C]// Proceedings of the 2011 International Symposium on Network Coding. Piscataway: IEEE Press, 2011: 1-6.

[3] WU W, MA R T B, LUI J C S. On incentivizing caching for P2P-VoD systems[C]// Proceedings of the 2012 IEEE Conference on Computer Communications Workshops. Piscataway: IEEE Press, 2012: 164-169.

[4] SU X, DHALIWAL S. Incentive mechanisms in peer-to-peer media streaming systems[C]// IWQOS 2004: Proceedings of the Twelfth IEEE International Workshop on Quality of Service. Piscataway: IEEE Press, 2004: 171-180.

[5] VISHNUMURTHY V, CHANDRAKUMAR S, SIRER E G. KARMA: a secure economic framework for peer-to-peer resource sharing [EB/OL]. [2013-03-11]. http://netecon.seas.harvard.edu/P2PEcon03.html/Papers/Vishnumurthy_03.pdf.

[6] ZHOU L, ZHENG R, WAN J. On delayed micropayment based incentive mechanism for peer-to-peer streaming media[J]. Computer Applications and Software, 2009, 26(1): 210-211. (周丽, 郑若艇, 万健. 基于延期微支付的 P2P 流媒体激励机制研究[J]. 计算机应用与软件, 2009, 26(1): 210-211.)

[7] HAUSHEER D, STILLER B. PeerMart: the technology for a distributed auction-based market for peer-to-peer services[C]// ICC 2005: Proceedings of the 2005 IEEE International Conference on Communications. Piscataway: IEEE Press, 2005, 3: 1583-1587.

[8] HAUSSHEER D, STILLER B. Decentralized auction-based pricing with PeerMart[C]// IM 2005: Proceedings of the 2005 9th IFIP/IEEE International Symposium on Integrated Network Management. Piscataway: IEEE Press, 2005: 381-394.

[9] LIU H T, HUANG Z X, BAI Y, *et al*. Auction incentive mechanism in P2P[C]// MUE 2007: Proceedings of the International Conference on Multimedia and Ubiquitous Engineering. Piscataway: IEEE Press, 2007: 941-945.

[10] GUO D, KWOK Y K. A new auction based approach to efficient P2P live streaming[C]// Proceedings of the 2011 17th IEEE International Conference on Parallel and Distributed Systems. Piscataway: IEEE Press, 2011: 573-580.

[11] ZHAO H. Research on key algorithms of P2P VoD system[D]. Chendu: Xihua University, 2008. (赵惠. P2P 流媒体点播系统的算法研究[D]. 成都: 西华大学, 2008.)

[12] ZUO F, ZHANG W. Selfish allocation avoidance for P2P file application: a game theoretic approach[C]// Proceedings of the 2012 18th IEEE International Conference on Networks. Piscataway: IEEE Press, 2012: 441-446.

[13] SUN W, XIA Q, XU Z, *et al*. A game theoretic resource allocation model based on extended second price sealed auction in grid computing[J]. Journal of Computers, 2012, 7(1): 65-75.

[14] ZENG W, XU Y. Peer-to-peer video on demand system with partitioned buffer-scheduling strategy[J]. Computer Engineering, 2010, 36(9): 90-93. (曾文烽, 许胤龙. 采用分区缓存调度策略的 P2P 点播系统[J]. 计算机工程, 2010, 36(9): 90-93.)


(上接第 3146 页)


[3] YI B, WANG Y, CHEN X, *et al*. Extracting hot topics from microblogging based on keywords detection and text clustering[J]. Applied Mechanics and Materials, 2013, 303-306: 2289-2293.

[4] CUI D. Research and improvement on *K*-means clustering algorithm[D]. Hefei: Anhui University, 2012: 1-5. (崔丹丹. *K*-means 聚类算法的研究与改进[D]. 合肥: 安徽大学, 2012: 1-5.)

[5] LI X. A new text clustering algorithm based on improved k_means[J]. Journal of Software, 2012, 7(1): 95-101.

[6] GUPTA N, SAXENA P C, GUPTA J P. Automatic generation of initial value *k* to apply *K*-means method for text documents clustering[J]. International Journal of Data Mining, Modelling and Management, 2011, 3(1): 18-41.

[7] ZHAO P, CAI Q. Research of Chinese text clustering algorithm based on HowNet[J]. Computer Engineering and Applications, 2007, 43(12): 162-163. (赵鹏, 蔡庆生, 一种基于《知网》的中文文本聚类算法的研究[J]. 计算机工程与应用, 2007, 43(12): 162-163.)

[8] ZHENG Y, SHU J, CHUN L, *et al*. A text hybrid clustering algorithm based on HowNet semantics[J]. Key Engineering Materials, 2011, 474-476: 2071-2078.

[9] ZHAO S, LIU T, LI S. A topical document clustering method[J]. Journal of Chain Information Processing, 2007, 21(2): 58-62. (赵世奇, 刘挺, 李生. 一种基于主题的文本聚类算法[J]. 中文信息学报, 2007, 21(2): 58-62.)

[10] YUAN X. A topic-based Web text clustering algorithm[J]. Journal of Chengdu University: Natural Science, 2010, 29(3): 249-252. (袁晓峰. 一种基于主题的 Web 文本聚类算法[J]. 成都大学学报: 自然科学版, 2010, 29(3): 249-252.)

[11] DONG J. Document clustering method based on LDA model[D]. Wuhan: Central China Normal University, 2012: 25-41. (董婧灵. 基于 LDA 模型的文本聚类研究[D]. 武汉: 华中师范大学, 2012: 25-41.)

[12] MAO J. Research of Chinese Web text clustering technology[D]. Xiamen: Xiamen University, 2007: 11-12. (茅剑. 中文 Web 文本聚类研究[D]. 厦门: 厦门大学, 2007: 11-12.)

[13] KWALE F M. A critical review of k means text clustering algorithm[J]. International Journal of Advanced Research in Computer Science, 2013, 4(9): 27-34.