综 述

文本聚类中文本表示和相似度计算研究综述

吴夙慧[1],成 颖[1],郑彦宁[2],潘云涛[2]

(1.南京大学 信息管理系,江苏 南京 210093;2.中国科学技术信息研究所,北京 100038)

摘 要:围绕文本聚类中的文本表示和相似度计算两个基本的问题,对目前学界提出的文本表示方法和相似度计算方法进行了分类和较为全面的综述,将文本表示模型分为向量空间模型、语言模型、后缀树模型、本体等,相似度计算方法分为基于向量空间模型的相似度计算,基于短语的相似度计算方法和基于本体的相似度计算方法。

关键词:文本聚类;文本表示;相似度计算

中图分类号:G350 **文献标识码**:A **文章编号**:1007-7634(2012)04-622-06

A Survey on Text Representation and Similarity Calculation in Text Clustering

WU Su-hui[1], *CHENG Ying*[1],*ZHENG Yan-ning*[2], *PAN Yun-tao*[2]

(1.*Department of Information Management,Nanjing University, Nanjing* 210093, *China;*
2. *Institute of Scientific and Technical Information of China, Beijing* 100038, *China)*

Abstract: The two basic problems of text clustering are text representation and similarity calculation. In this paper, We classified the different text representation models and the methods of similarity calculation and summarized them detailedly. This paper classified the present text representation models as VSM, language model, suffix tree model and ontology, classified the methods of similarity calculation as three categories, including VSM-based method, phrase-based method and ontology-based method.

Key Words: text clustering; text representation; similarity calculation

1 文本表示

文本表示是信息检索、文本挖掘、自然语言处理等领域的基础,其目的是将非结构化原始文本转换为计算机能够处理的数据形式。至今,学术界已经提出了多种文本表示模型,传统的文本表示模型有向量空间模型,语言模型等,而随着语义网和本体研究的深入,近几年来研究较多的是基于本体的文本表示模型。

1.1 向量空间模型

目前,文本聚类研究大多采用向量空间模型(VSM)进行文本表示[1-2],许多实际运行的信息检索系统也采用了这样的模型。该模型由Salton于1975年提出[3],其基本思想是将文本表示成一个向量,向量的每一维表示文本的一个特征,该特征通常是一个字或词。使用VSM进行文本表示,需要进行的工作包括分词、停用词处理、词根处理以及权重计算等,然后文本集D中的任一文本d_j都可以表示成形如$(W_{1j},W_{2j},\cdots\cdots,W_{nj})$的向量,其中,$W_{ij}$表示文本$d_j$中的词的权重。权重计算的方法主要有TFIDF函数、布尔函数、频度函数等。其中使用较多的是由Salton[4]提出的TFIDF函数。

向量空间模型以其简单的表示方法,良好的文本表示效果得到了学界的青睐。不过,由于向量空

收稿日期:2011-10-12

基金项目:国家社科基金项目(10CTQ027);教育部人文社会科学研究规划基金项目(07JA870006);中国科学技术信息研究所合作研究项目

作者简介:吴夙慧(1988-),男,安徽安庆人,硕士生,主要从事自然语言处理研究.

间模型在文本表示中采用独立性假设，即认为词与词之间是相互独立的，割裂了文本原有的语义关系，舍弃了词与词之间大量的联系，因而该模型的局限性也比较大。

针对VSM的不足，学界提出了语言模型、后缀树模型以及本体等文本表示方法以更多地保留文本语境，以增强特征项之间的语义联系。

1.2 语言模型

语言模型最早由Ponte和Croft[5]在1998年应用到文本检索领域。语言模型是描述词，句子等语言基本单位的分布函数。目前，研究较多的是统计语言模型，该模型通过前期大规模语料库的学习和统计得到真实语料中的语言知识，如词与词之间的共现关系、上下位关系等。利用这些知识可以计算出一个特征项出现在文本中的概率，最终将整个文本表示为所有特征项的概率分布。由于在一定程度上描述了语义层面的知识，因此语言模型可以更好地进行文本表示，这是向量空间模型所不及的。

目前，基于语言模型的文本聚类研究主要由中国学者完成。2006年，Zhang XD，Zhou XH等[6-7]利用语义平滑(Semantic Smoothing)的概念，采用短语作为文本表示单位，解决了文本聚类中上下文无关的问题，同时削弱了常用高频词在文本表示中的权重，更好地反映了文本内容。作者实现了凝聚和K-means两种不同聚类算法，实验结果表明，当数据稀疏时，凝聚算法在语义平滑模型中得到了更好的聚类效果，而划分算法在数据集较小时则具有更佳的聚类效果。2008年，Wen J和Li ZJ[8]在语义平滑和混合概率模型的基础上提出了混合语言模型，并实现了基于混合语言模型的期望最大化聚类算法，算法采用语义平滑模型表示文本，再利用混合概率模型计算相似度。

与VSM的广泛应用相比，国内外关于语言模型的研究还主要集中在信息检索，词聚类，文本自动分类等领域，基于语言模型的文本聚类研究还较少。不过，由于语言模型的优势，其在文本聚类领域应该大有可为。

1.3 后缀树模型

后缀树是一种用于字符串处理的数据结构，最早由Weiner[9]提出。1998年，Oren Zamir等[10]进行了基于后缀树模型的聚类研究。与VSM不同，后缀树模型将文本视为短语的集合，相比单词，短语显然具有更多语义上的关联和上下文关系，因此能更好地表示文本的特征。所谓的字符串的后缀是指从字符串中的某一个词开始到最后一个词所构成的字符串。对于一个长度为m的字符串而言，其拥有m个后缀，而后缀树就是一棵由这m个后缀所组成的树。对于多个字符串，就需要引入广义后缀树的概念，即将多棵后缀树连接起来，置于同一个根节点之下。图1是一个介绍广义后缀树模型的的经典例子，是一棵由“cat ate cheese”，“mouse ate cheese too”，“cat ate mouse too”三个短语组成的广义后缀树。

1998年，Zamir和Etzioni[11]实现了基于后缀树的STC算法(Suffix Tree Clustering)。Wang等[12]对后缀树的数据结构进行了改进，提出一种基于N元语法的后缀树模型，传统的后缀树由文本中所有出现的短语构成，造成后缀树十分庞大，使得树中存在许多的冗余信息。而基于N元语法的后缀树限制了短语的长度和出现频率，仅将出现次数较多且长度小于N的短语并入后缀树中，作者将这样的后缀树模型应用于Web检索结果的聚类，在聚类速度和聚类效果上都有所提升，尤其在聚类标签的提取上更有着不俗的表现。史庆伟等[13]结合Chameleon聚类算法改进了STC算法，杜红斌等[14]提出了一种改进的后缀树聚类算法，对STC算法的几个问题进行了改进，提出了一种新的基类合并的相似度计算公式，并且采用信息增益的方法提出了新的聚类标识的提取算法。

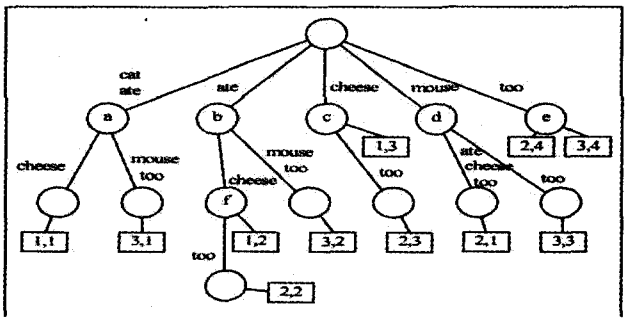

图1 “cat ate cheese”，“mouse ate cheese too”，“cat ate mouse too”后缀树[12]

1.4 本体

近几年，随着语义网和本体研究的深入，本体的引入也为文本表示提供了新的可能。本体是特定领域客观存在的概念及概念之间关系的描述。在文本表示中，可以将一个文本分解为多个概念，利用本体得到概念之间的语义关系。基于本体的文本表示方

法揭示了文本语义层面的知识,无疑会带来更好的文本表示效果。早期的研究主要来自生物学领域,随着对生物过程、分子功能和细胞组建三个数据库的成功整合,基因本体(Gene Ontology)得以建立,生物学领域的许多工作都基于此完成,其中,大规模数据的聚类是其非常重要的应用,2004年至今已有多篇该方向上的研究[15-17]。

基于本体的文本聚类领域的研究,国外较早的有Stabbs和Hotho[18]的工作。2006年,Wang[19]提出了基于本体的支持向量机聚类算法,并通过实验证明了算法的效率。Karoui[20]等采用本体学习的方法提出了一种基于上下文的层次聚类COCE算法,算法运用词语之间的上下文关系,挖掘出了文本的深层语义知识,比起传统的聚类算法更为优秀。2008年,Jing等[21]采用基于本体的互信息测度来计算特征项之间的相似度,利用WordNet本体得到两个特征项之间的距离,再通过计算特征项与文档之间的互信息值来得到特征项在该文档中的权值,将文档表示成特征项的权值向量,最终通过计算向量的距离得到文档之间的相似度从而完成文本聚类。2009年,Song等[22]提出了基于本体的遗传聚类算法,算法利用WordNet本体进行文本表示,并提出一种本体中的概念相似度计算方法,最终采用遗传算法进行聚类。此外,本体也被应用到一些传统的聚类算法,如遗传算法,模糊聚类算法,基于密度的算法等。

在国内,2006年罗娜等[23]提出了可以用本体语义改进聚类效果的设想。2008年,谢红薇等[24]在VSM中应用本体的概念,将文本中每个特征项与本体匹配,进而调整特征项的权值,得到新的特征向量,改进了VSM缺乏语义知识的不足。2010年,朱会峰等[25]采用WordNet作为本体实现了基于本体的聚类算法,并与其他聚类算法进行了比较,实验结果证明,WordNet增强了文本表示的效果,大大降低了文本向量的维度,得到了更好的聚类结果。

语言模型、后缀树模型以及本体融入的语义知识,更为真实地反映了文本的内容特征,实证研究也证实其可以获得更好的聚类效果[26-27],因此,它们的应用是文本表示的一个重要方向。

2 相似度计算

文本聚类是将文本对象划分为若干簇,使得簇内的文本尽可能相似,簇间的文本尽量相异。判断文本对象间相似和相异的程度需要一个量化的尺度,这就是相似度计算。相似度计算依赖于文本表示模型,因此,文本表示方法的差异通常也意味着相似度计算方法的不同。

2.1 基于向量空间模型的相似度计算

向量空间模型是将文本表示成向量空间中的一个点,较为简单的相似度计算方法是通过数学方法计算点与点之间的距离,并以此作为文本相似度。数学上应用较广泛的距离计算方法有欧氏距离,街区距离,幂距离等[28]。

对于文本集中的两个文本向量 $\vec{d}_i$ ($W_{1i}, W_{2i}, \cdots, W_{ni}$), $\vec{d}_j$ ($W_{1j}, W_{2j}, \cdots, W_{nj}$),它们之间的距离计算公式为欧氏距离、街区距离、幂距离等。

除了距离相似度之外,学界还提出了基于K最近邻集(KNN)的相似度计算方法。

Chris Ding和He[29]在计算相似度时考虑邻近点的特性,引入了K最近邻的概念,即使用邻近点作为相似度的评判标准。K最近邻集是指空间中离某点最近的K个点的集合。基于K最近邻,Chris Ding等提出了K-means-CP算法。与K-means算法相比,K-means-CP算法不再以单个点作为聚类对象,而以点的近邻集作为聚类单位,从而得到了更好的聚类效果。

2006年,黄建鹏,陆力强[30]在K最近邻概念的基础上提出了共享K最近邻的概念。所谓的共享K最近邻的概念,即空间中两个点x,y互为K最近邻,记为$x \in SNN^k(y)$。在共享K最近邻的概念上定义了两个点和两个点集之间的第一连接度及其计算方法,即若空间中两个点是共享K最近邻,称两个点第一连接,第一连接度的值为1,否则取0,计算公式为:

$$LINK^k(x,y)=\begin{cases}1 & x \in SNN^k(y) \\ 0 & \text{其他}\end{cases} \quad (1)$$

对于两个不相交点集A,B,第一连接度的值计算公式为:

$$\sum_{\substack{x \in A \\ y \in B}} LINK^k(x,y) \quad (2)$$

如果空间中两点x,y有一个点属于另一个点的K最近邻集,则称两个点第二连接,第二连接度的值为1,否则取0,计算公式为:

$$link^k(x,y)=\begin{cases}1 & x \in KNN^k(y) \text{或} y \in KNN^k(x) \\ 0 & \text{其他}\end{cases} \quad (3)$$

对于两个不交点集A,B,第二连接度的值计算

公式为：

$$\sum_{\substack{x\in A\\ y\in B}} link^{k}(x,y) \tag{4}$$

作者通过数学方法证明：可以用第一连接度和第二连接度之比作为相似度，实验结果表明，该相似度方法可以取得更优的聚类效果。

2.2 基于短语的相似度计算方法

为了在文本表示中增强文本的语义联系，许多文本表示方法采用了短语作为特征项，其相似度计算方法也与向量空间模型存在很多的不同。

Hammouda 和 Kamel[31-33]提出了一种基于短语的文本相似度计算方法，其基本思想是采用两个文本之间相交的短语占两个文本短语并集的比例作为文本相似度。对于相交短语需要综合考虑短语的个数、长度、在两个文本中的出现频度以及语义重要性等指标，进行加权后得到文本的相似度。作者据此对Web文本进行了聚类，实验证明如果使用单词相似度和短语相似度相结合的方法，会得到更好的聚类结果，计算公式为：

文本相似度=α短语相似度+(1-α)单词相似度 (5)

其中，单词相似度的计算方法与上面短语相似度的计算方法类似，实验结果表明，当$\alpha\in[0.6,0.8]$时，聚类结果最优。

在后缀树模型中，文本被表示为短语的集合。在一棵后缀树中，中间节点表示两个词串所共有的短语，因而可以用两个文本的后缀树所共有的中间节点数来定义文本相似度。

2005年，Eissen等[34]提出了3种基于后缀树的相似度计算方法。给定文本d^+,d^-，将文本集中的所有文本构成一颗广义后缀树T，记T中的所有边的集合为E。

第一种相似度记为φ ST，即后缀树T中d^+,d^-边的交集除以d^+,d^-边的并集。计算公式为：

$$\varphi_{ST}=\frac{|E^+\cap E^-|}{|E^+\cup E^-|} \tag{6}$$

第二种相似度记为φ STF，考虑了短语在文本中的出现频度。计算公式如下：

$$\varphi_{STF}=\frac{1}{|E|}\sum_{e\in E}\frac{\min\{n^+(e),n^-(e)\}}{\max\{n^+(e),n^-(e)\}} \tag{7}$$

其中，$n^+(e)$和$n^-(e)$代表后缀树T中的边e在文本d^+,d^-中的出现次数。

第三种相似度记为φ STIDF，在公式(6)、(7)的基础上引入了IDF的概念，即考虑了短语在整个文本集中的出现频率。计算公式为：

$$\varphi_{STFIDF}=\frac{1}{|E|}\sum_{e\in E}\frac{\min\{n^+(e),n^-(e)\}}{\max\{n^+(e),n^-(e)\}}\cdot IDF(e) \tag{8}$$

2008年，Chim H和Deng XT[35]提出了一种新的基于后缀树模型的相似度计算方法，算法采用了一种类似于TFIDF的思想，即对文本后缀树中的所有中间节点和叶子节点，计算其TF和DF值，节点的TF值定义节点在该文本中出现的频率，同样地，DF值定义为在整个文本集中，出现该节点的文本数量。使用TF值和DF值对节点加权，得到文本的向量表示，这样就可以运用类似于向量空间模型的方法计算相似度。作者运用凝聚层次聚类算法对这种相似度计算方法和基于单词的TFIDF计算方法进行了比较，证明了新算法的优越性。

2010年，Yang[36]等将加权的思想应用于后缀树模型，在构建后缀时考虑了短语的出现位置，认为出现在题名、关键词等位置的短语相比正文中的短语更为重要，赋予不同的权重，在计算相似度时将权重考虑进来。

2.3 基于本体的相似度计算方法

在前面提到过，本体作为一种优秀的工具，将会被越来越多地应用到文本聚类中，而基于本体的相似度计算方法也已比较成熟。以本体的观点，文本被看作为由本体构成的集合，而本体又是由概念和关系组成的，所以最终就可以将文本之间的相似度转化为概念之间的相似度。

国外对于这一领域的研究很多，Zhang等[37]将国外相似度计算方法总结为基于距离的相似度计算，基于信息内容的相似度计算，基于特征的相似度计算三大类。孙海霞等[38]在Zhang的基础上将基于本体的语义相似度计算方法分为四大类，补充了混合式相似度计算方法，并指出了影响本体中语义相似度计算的五个因素。

在国内，也有许多学者致力于计算方法的改进，吕刚与郑诚[39]提出可以根据本体中的概念的深度和密度的不同对概念进行加权，从而得到更加精确，更反映事实的相似度。胡哲等[40]提出在计算概念相似度时需要更多地考虑概念之间的上下位关系。

3 结 语

本文围绕着文本聚类中文本表示和相似度计算

对国内外相关领域的研究进行了综述。可以发现经过多年的研究,无论是传统的向量空间模型还是后缀树、本体等文本表示模型都已经可以较为成熟的应用于文本聚类,并取得了不错的聚类效果。近年来文本聚类的相关研究依然大量集中在算法思想的改进上,然而这些改进并没有使聚类效果获得很大地提升。因而要从算法上较大地提高聚类效果已经十分困难,因此需要在其他方向另辟蹊径。

基于语料库的文本聚类研究就是一个不错的研究方向。在自然语言处理的很多研究领域,利用大规模语料库进行前期的训练已经成为改进算法效果的重要手段,而大多数聚类算法都属于无监督学习的算法,在算法的前期并没有进行任何的训练和学习。因而可以尝试在聚类前先通过语料库进行训练,从而得到一些先期的知识,应该对改进聚类效果有所帮助。1990年贝尔实验室的Hindle[41]提出了基于语料库的英文名词聚类算法可以为文本聚类研究提供借鉴。Hindle认为,如果一个名词可以与一个动词搭配,则它们之间存在一定的互信息量,两个名词之间如果拥有相同的修饰动词,则之间也存在一定的互信息量,最终就可以运用聚类算法完成单词的聚类。如果把聚类的粒度从单词扩大到文本,该算法则完全可以扩展到文本聚类。

此外,一些学者利用词语或句子之间上下文以及共现关系建立词语或句子网络,这些词句网络在许多特性上类似于链接网络和社会网络,因而链接分析方法和社会网络分析方法这样一些较为成熟的研究方法也被应用到词句关系网络上,为文本挖掘和自然语言处理的研究提供了一个新的方向。Wan和Yang[42],Wan和Xiao[43]分别利用链接分析算法——HITS算法的思想实现了多文档自动摘要,获得了不俗的实验效果。这样的思想同样可以嫁接到文本聚类的研究中,通过构建文本网络,利用PageRank和HITS等算法在文本网络上挖掘以获得一些语义层面上的知识,这样的方法也为文本聚类效果和聚类标签质量的提高提供了一种新的思路。

参考文献

1 尉景辉,何丕廉,孙越恒. 基于K-Means的文本层次聚类算法研究[J].计算机应用, 2005,25(10):2323-2324.

2 姚清耘,刘功申,李 翔. 基于向量空间模型的文本聚类算法[J].计算机工程,2008,34(18):39-41,44.

3 Salton G, Wong A,Yang C S. A Vector Space Model for Automatic Indexing[J]. Communication of the ACM, 1975,18(11): 613-620.

4 Salton G, Clement T Y. On the Construction of Effective Vocabularies for Information Retrieval[EB/OL]. http://dl.acm.org/citation.cfm?id=951766,2011-02-04.

5 Ponte J M, Croft W B. A Language Modeling Approach to Information Retrieval[EB/OL]. http://dl.acm.org/citation.cfm?id=291008,2011,02-04.

6 Zhang X,Zhou X,Hu X. Semantic Smoothing for Model-based Document Clustering[EB/OL]. http://www.cis.drexel.edu/faculty/thu/research-papers/ICDM2006_Clustering.pdf, 2011, 02-04.

7 Zhou X,Zhang X,Hu X. Semantic Smoothing of Document Models for Agglomerative Clustering[EB/OL]. http://ijcai.science.unitn.it/Past_Proceedings/IJCAI-2007/PDF/IJCAI07-470.pdf,2011-02-04.

8 Wen J, Li Z. Research on Mixture Language Model-based Document Clustering[EB/OL]. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4664755,2011-02-04.

9 Weiner P. Linear pattern matching algorithms[EB/OL].http://airelles.i3s.unice.fr/files/Weiner.pdf,2011-02-06.

10 Zamir O , Etzioni O. Web Document Clustering[EB/OL]. http://dl.acm.org/citation.cfm?id=290956,2011-02-06.

11 Zamir O, Etzioni O. Web Document Clustering: a Feasibility Demonstration[EB/OL]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.122.6420&rep=rep1&type=pdf,2011-02-06.

12 Wang JZ, Mo YJ, Huang BX, Wen J, He L. Web Search Results Clustering Based on a Novel Suffix Tree Structure[EB/OL]. http://www.springerlink.com/content/h56785141260hh02/,2011-02-06.

13 史庆伟,赵政,朝柯. 一种基于后缀树的中文网页层次聚类方法[J].辽宁工程技术大学学报,2006,25(6):890-892.

14 杜红斌,夏克文,刘南平,吴 涛.一种改进的基于广义后缀树的文本聚类算法[J].信息与控制,2009,38(3):331-336.

15 Adryan B, Schuh R. Gene-Ontology-based Clustering of Gene Expression Data[J]. BIOINFORMATICS,2004,20(16): 2851-2852.

16 Speer N, Spieth C, Zell A. A Memetic Clustering Algorithm for the Functional Partition of Genes Based on the Gene Ontology[EB/OL].http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.6448&rep=rep1&type=pdf,2011-02-06.

17 Grotkjaer T, Winther O, Regenberg B, Nielsen J, Hansen LK. Robust multi-scale Clustering of Large DNA Microarray Datasets with the Consensus Algorithm[J].BIOINFORMATICS, 2006,22(1):58-67.

18 Hotho A, Maedche A, Staab S. Ontology-based Text Document Clustering[EB/OL]. http://www.aifb.kit.edu/images/2/2b/2002_19_Hotho_Text_Clustering_1.pdf,2011-02-06.

19 Wang D, Wang R, Wang Y, Qiu D. Clustering by SVM based on Ontology[EB/OL]. http://ieeexplore.ieee.org/xpls/abs_all.

jsp?arnumber=85677,2011-02-07.

20 Karoui L, Aufaure MA, Bennacer N. Context-based Hierarchical Clustering for the Ontology Learning[EB/OL]. http://dl.acm.org/citation.cfm?id=1249167,2011-02-07.

21 Jing LP,Zhou LX,NG MK,Huang ZX. Ontology-based Distance Measure for Text Clustering[EB/OL]. http://www.siam.org/meetings/sdm06/workproceed/Text% 20Mining/jing1.pdf, 2011-02-07.

22 Song W,Li CH,Park SC. Genetic Algorithm For Text Clustering Using Ontology And Evaluating The Validity Of Various Semantic Similarity Measures[J]. Expert Systems with Applications. 2009,(36): 9095-9104.

23 罗 娜,左万利,袁福宇,张靖波,张慧杰.使用本体语义提高文本聚类[J].东南大学学报(英文版),2006,22(3):370-373.

24 谢红薇,颜小林,余雪丽.基于本体的WEB页面聚类研究[J].计算机科学,2008,35(9):153-155.

25 朱会峰,左万利,赫枫龄,彭 涛,纪文彦.一种基于本体的文本聚类方法[J].吉林大学学报(理学版),2010,48(2): 277-283.

26 Huang RZ,Lam W. An Active Learning Framework for Semi-supervised Document Clustering with Language Modeling[J]. Data & Knowledge Engineering, 2009,68(1):49-67.

27 Han S, Lee SG, Kim KH,Choi CJ, Kim YH, Hwang KS. CLAGen: A Tool for Clustering and Annotating Gene Sequences using a Suffix Tree Algorithm[J]. Biosystems, 2006,84(3): 175-182.

28 Soman KP, Diwakar,Ajay. 数据挖掘基础教程[M].北京:机械工业出版社,2009:215-216.

29 Ding C, He X. K-Nearest-Neighbor in Data Clustering: Incorporating Local Information into Global Optimization[EB/OL]. https://mailserver.di.unipi.it/ricerca/proceedings/AppliedComputing04/Papers/T09P13.pdf,2011-02-07.

30 黄建鹏,陆力强. 一种新的相似度标准及其相关的聚类算法[J].复旦学报(自然科学版), 2006,45(2):177-184.

31 Hammouda KM, Kamel MS. Phrase-based Document Similarity based on an Index Graph Model[EB/OL]. http://watnow.uwaterloo.ca/pub/hammouda/hammouda_icdm02.pdf,2011-02-07.

32 Hammouda KM, Kamel MS. Incremental Document Clustering using Cluster Similarity Histograms[EB/OL]. http://watnow.uwaterloo.ca/pub/hammouda/wi03.pdf,2011-02-07.

33 Hammouda KM, Kamel MS. Efficient phrase-based Document Indexing for Web Document Clustering[J]. IEEE Transactions On Knowledge And Data Engineering, 2004,16(10):1279-1296.

34 Eissen SM,Stein B, Potthast M. The Suffix Tree Document Model Revisited[EB/OL]. http://i-know.tugraz.at/wp-content/uploads/2008/11/29_the-suffix-tree-document-model-revisited.pdf,2011-02-07.

35 Chim H, Deng X. Efficient Phrase-Based Document Similarity for Clustering[J].IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2008, 20(9):1217-1228.

36 Yang RL, Zhu QS, Xia YN. Weighted Suffix Tree Document Model for Web Documents Clustering[EB/OL]. Weighted Suffix Tree Document Model for Web Documents Clustering,2011-02-08.

37 Zhang XD, Jing LP, Hu XH, NG MK, Xia JL, Zhou XH. Medical Document Clustering Using Ontology-Based Term Similarity Measures[J]. International Journal of Data Warehousing & Mining, 2008,4(1): 62-73.

38 孙海霞,钱 庆,成 颖. 基于本体的语义相似度计算方法研究综述[J]. 现代图书情报技术, 2010,(1):51-56.

39 吕 刚,郑 诚. 基于加权的本体相似度计算方法[J].计算机工程与设计,2010,31(5):1093-1095.

40 胡 哲,郑 诚. 改进的概念语义相似度计算[J].计算机工程与设计,2010,31(5):1121-1124.

41 Hindle D. Noun Classification from Predicate-argument Structures[EB/OL]. http://acl.ldc.upenn.edu/P/P90/P90-1034.pdf,2011-02-08.

42 Wan XJ, Yang JW. Multi-Document Summarization Using Cluster-Based Link Analysis[EB/OL]. http://clair.si.umich.edu/si767/papers/Week08/Summarization/p299-wan.pdf,2011-02-08.

43 Wan XJ, Xiao JG. Graph-Based Multi-Modality Learning for Topic-Focused Multi-Document Summarization[EB/OL].http://robolab.cse.unsw.edu.au/conferences/IJCAI-2009/08/IJCAI09-265.pdf,2011-02-08.

(责任编辑:刘凤琴)