●章成志　张庆国　师庆辉

基于主题聚类的主题数字图书馆构建*

摘　要　基于主题聚类的主题数字图书馆是针对某一个特定的主题,获取与该主题相关的数字资源集合(本文以文本资源为研究对象),然后再依据主题聚类算法,对该主题的信息资源集合进行聚类,生成可供用户浏览的多层次结构导航,结合全文检索实现基于主题聚类的主题数字图书馆系统。主题数字图书馆系统主要包括主题采集模块、主题聚类模块和数据集成模块,构建过程中主要涉及主题提取、主题聚类以及聚类结果描述等三类关键技术。表 2。图 1。参考文献 20。

关键词　数字图书馆　主题聚类　主题抽取　文本聚类

分类号　G254.2

ABSTRACT　Subject digital library based on subject clustering is a subject digital library system focused on a particular subject, collecting subject-related digital resources (mainly text resources in this article), clustering information resource subsets by subject clustering algorithms, generating hierarchically structured navigation for users and integrating full-text retrieval. The system should include a subject acquisition module, a subject clustering module and a data integration module. In the construction of such a system, we can use three key technologies of subject extraction, subject clustering and subject result description. 2 tabs. 2 figs. 20 refs.

KEY WORDS　Digital library.　Subject clusetering.　Subject extraction.　Text clustering.

CLASS NUMBER　G254.2

1　引言

我们正处于“信息爆炸”的时代,用户迫切希望获得信息噪声尽量少、乃至没有噪声的信息服务。传统信息组织方法在有效开发利用信息资源方面发挥重要作用,但在数字图书馆建设当中,由于主题法、分类法或分类主题一体化方法要依赖于大量专业人员的参与,使得绝大多数机构面临人力和财力不足的困境。传统信息组织方法置身于互联网海量数据环境中,无法充分及时地为满足用户信息需求提供便利,这迫使人们去寻找解决海量数据自动化处理的方法和技术。文本自动聚类是一种典型的无指导学习方法。由于一般的文本聚类方法直接以词语作为特征,文本向量空间的维度可能会达到数万以上,这使得应用服务过程中面临高维数据计算问题。同时,由于缺乏充分的主题控制和语义理解机制,会导致聚类技术在应用服务过程中出现大量信息噪声。传统的聚类技术直接用于文档聚类,存在的突出问题是传统算法只对对象进行聚类,不负责对聚类后生成的类簇进行概念描述和解释。

针对以上问题,传统信息组织方法与数据挖掘方法的有效融合已是当务之急。传统信息组织方法中的主题法与数据挖掘、机器学习中的聚类方法的结合,使得主题聚类方法应运而生。主题聚类主要通过对信息集合进行主题分析与提取,获得能表达其主题的关键词或主题词集合后,再对该集合进行聚类,最后得到基于主题的聚类结果描述。

目前,有关主题法、分类法、主题分类一体化以及索引法的研究已非常深入,聚类方法研究也比较细致,但对主题法与聚类方法的融合体,即主题聚类研究,仅有很少学者涉及,如马张华等人提出基于控制词集的中文信息动态自动聚类技术,从词汇控制角度对动态聚类进行了相对全面的研究[1];Kang、Chang 等分别进行基于关键词聚类的文本聚类研究[2-3];孙学刚等人采用二次特征提取和聚类方法,对 Web 文档按照主题进行聚类[4];Zhao & George 提出主题驱动的文本聚类方法[5];赵世奇等人通过文本主题元素的索引完成聚类[6],等等。此外,与主题聚类研究相关的系统有:基于 SOM 神经网络方法的数字图书

* 本研究受“十一五”国家科技支撑计划重点项目“科技文献信息服务系统关键技术研究及应用示范”子课题(2006BAH03B02、2006BAH03B04)、南京理工大学青年科研扶持基金项目“基于机器学习方法的领域本体学习研究”(JGQN0701)和南京理工大学科研启动基金项目“主题聚类关键技术研究”(AB41123)资助。

馆系统 SOMLib[7]、基于聚类的文档浏览系统 Scatter/Gather[8]等。这些研究基本上只涉及主题聚类研究的一部分,缺少主题聚类的系统化研究,也缺乏对主题聚类中各个步骤相互影响的全面研究。

2　主题聚类基本原理

主题聚类(或称主题聚类一体化)信息组织方法,是融合信息组织方法中的主题法与数据挖掘中的聚类方法形成的一种特殊的知识组织方法,可作为信息资源的一种组织模式。它通过对聚类对象进行主题分析和提取,将聚类对象转换为基于主题的表示形式,达到降低特征空间维度和进行主题控制与语义理解的目的,然后以主题表示为基础进行对象的聚类,最后得到基于主题的聚类结果描述。

对主题聚类方法进行进一步分析可知,该方法具有三方面的优势。首先,主题聚类以主题分析、主题提取和描述为基础,可以发挥主题法在组织信息方面的优势,对聚类特征进行主题或语义控制,提高信息服务的质量。其次,主题聚类在聚类对象的主题提取基础上进行,通过主题提取可以对聚类对象进行维度约简,从而避免高维数据计算问题,缩短信息服务响应时间。最后,主题聚类方法还可对聚类结果进行基于主题的描述,提高聚类结果的可读性与可理解性。

主题聚类一般包括主题提取与样本聚类以及聚类结果描述三部分,主题聚类过程一般包括六个步骤:①对聚类样本进行词法、句法分析,获得聚类对象主题特征;②根据标引模型对聚类样本进行主题提取,并进行提取性能的评估;③依据文本表示模型,将样本主题提取结果映射到样本特征空间;④利用相似度计算模型计算聚类样本间的相似度;⑤使用聚类算法或模型对聚类样本进行聚类,并进行聚类性能的评估;⑥通过聚类描述算法获得聚类结果的概念描述,并对描述性能进行评估。

文本聚类是当前研究热点之一。提高聚类质量与实用化程度、提高聚类结果描述的可理解性等是文本聚类迫切需要解决的问题。本文从主题角度出发,提出主题聚类方法,并将主题聚类方法用于主题数字图书馆的构建,进行主题聚类应用方面的探索。

3　基于主题聚类的主题数字图书馆构建

主题数字图书馆是一种基于主题特征的专业领域数字图书馆,是针对某一特定主题的信息资源组织形式,其核心问题是主题资源的获取与主题聚类。与传统的文本聚类不同的是,基于主题聚类的主题数字图书馆针对某一个特定的主题,获取与该主题相关的数字资源集合(本文以文本资源为研究对象),然后再依据主题聚类算法,对该主题的信息资源集合进行聚类,生成可供用户浏览的多层次结构导航,结合全文检索实现基于主题聚类的主题数字图书馆系统。下文给出主题数字图书馆系统的总体设计方法,重点描述构建过程中的关键技术,即:主题提取、主题聚类以及聚类结果描述技术。

3.1　主题数字图书馆系统总体设计

构建主题数字图书馆系统的总体目标为:根据现有的信息资源,生成某一特定主题的数字图书馆系统,为该主题应用和研究提供专业的、有深度的信息资源采集、存储、聚类浏览和全文检索等服务。另外,在特定主题内部采用自动聚类技术发现产生内部聚合度较大的文件集合,可以实现特定领域在一段时间内热点的发现和相关文献的整合功能。

主题数字图书馆设计的基本思路为:利用现有的信息资源,生成某一主题的专题数据库,然后对该主题数据库进行基于主题的聚类,将生成的聚类描述作为该专题的子主题,并进一步生成主题导航,结合全文搜索功能,为用户提供主题浏览和检索服务。

主题数字图书馆主要包括三个模块,分别为主题采集模块、主题聚类模块、数据集成模块(见图 1)。

(1)主题采集模块。针对特定主题,从大规模科技文献集合(本文利用 CNKI 期刊数据库资源[9])提取与该主题相关的文献,形成此主题的专题库。利用规则(如依据文章的关键词进行采集)与统计方法[10]结合的文本分类方法,获取特定领域科技文献子集。

(2)主题聚类模块。在主题提取基础上对特定专题库进行自动聚类,给出聚类结果(即类簇)的描述标签,类别描述至多到两层目录。一篇文章可以属于多个子目录,引导用户在更确切的范围内,依靠子栏目导航功能,查找自己想要的文献内容。文本聚类模块涉及到主题提取、主题聚类以及聚类结果描述三个关键技术。

(3)数据集成模块。对特定专题库,通过主题聚类生成结果描述,生成供用户导航的分类层次结构,结合全文检索实现基于主题聚类的主题数字图书馆系统,便于用户进行资源的主题浏览和检索。

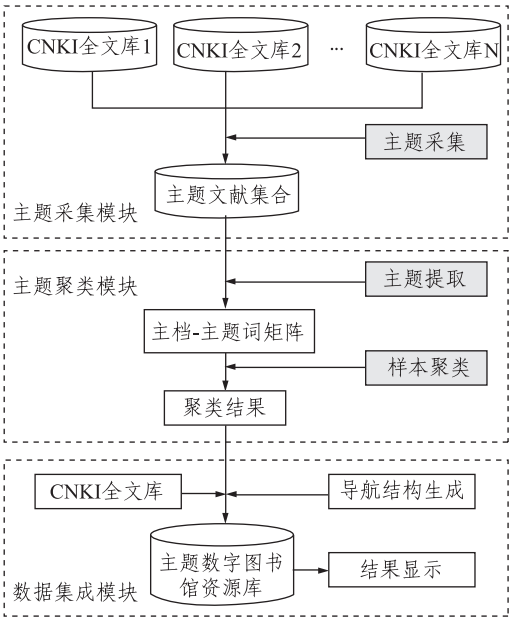

图 1　主题数字图书馆设计框架图

3.2　主题数字图书馆系统构建关键技术描述

如上所述,主题数字图书馆构建过程中涉及主题采集、文本聚类、数据集成等关键技术。由于自动分类技术相对成熟,因此本文对自动分类技术不作详细描述,而重点描述主题数字图书馆构建过程中的主题提取、聚类及聚类描述技术。

(1)主题提取

广义上的主题提取是指从文本中提取能表达文本主题的词语,而狭义的主题提取则是指提取能表达文本主题的主题词。本文所指的主题提取是广义的主题提取,提取的是文本的关键词。关键词自动提取方法有:统计方法[11]、语言学方法[12]、机器学习方法[13]及混合方法等[14]。

本文采用的主题提取方法是基于特征组合的方法[18]。该方法构造了一个大规模的关键词词典,计算候选词语 t 的 TF × IDF 值、关键词词频 *KeyFreq*(*t*)、词语直径 *Diameter*(*t*)、词语长度 *Length*(*t*)、词语首次出现位置 *FirstLoc*(*t*)、词语分布偏差 *Deviation*(*t*)等特征,然后使用乘算子来组合各个特征,计算各个特征共同作用下的候选词权重,计算公式如下:

$$Weight(t) = TF \times IDF \times KeyFreq(t) \times Diameter(t) \times FirstLoc(t) \times Length(t) \times Deviation(t) \quad (1)$$

给定关键词选取数目 K,候选关键词集合中前 K 个权重最大词汇就认为是关键词。本文给出煤炭类文献的关键词提取结果样例如下(K = 20):

工作面:4827,支架:4681,机采:3244,综采:2928,滞煤:2670,综采设备:2035,影响生产:1899,绞车:1820,片帮:1783,槽帮:1751,割煤:1575,顶梁:1481,综采工作面:1060,机组:944,运架:932,运输:883,盘区:874,上顺槽:855,井下生产:847,绳速:838

实验表明,基于特征组合方法同基于 Bayes 和 KNN 等机器学习方法性能相当。该方法的详细描述可参见文献[15]。

(2)样本聚类

聚类方法的研究发展至今,已经形成许多成熟的算法,并广泛应用到多个领域。各种聚类方法原则上都可以用在文本聚类上。常用于文本聚类的方法主要有两类,即基于系统树状图的等级聚类方法和基于平面划分的动态聚类方法。

本文以 K-Means 聚类算法为基础,在考虑样本本身对聚类的影响,即考虑样本权重的情况下,K-Means 聚类算法在准则函数收敛时结束聚类,其中准则函数如公式(2)所示:

$$J = \sum_{i=1}^{K}\sum_{j=1}^{m_i}(Sim(\vec{d}_j, \vec{c}_i)) \quad (2)$$

其中,J 为凝聚度,可用来度量聚类效果;K 为类簇的总数目;m_i 是类簇 i 中的成员总数;$\vec{d}_j$ 为类簇 i 中的第 j 个成员;$\vec{c}_i$ 为类簇 i 的中心向量,通过式(3)计算得到:

$$\vec{c}_i = \sum_{j=1}^{m_i}(w_j \cdot \vec{d}_j) \quad (3)$$

其中:w_j 为聚类样本 i 的权重,$\sum_{j=1}^{m_i} w_j = 1$,本文通过计算论文的 PageRank 值,将其作为样本的权重。论文 PageRank 值计算方法参见文献[16]。$Sim(\vec{d}_j, \vec{c}_i)$ 为文本 $\vec{d}_j$ 与类簇中心 $\vec{c}_i$ 点的相似度,文本和类簇的中心向量都是利用向量空间模型表示,经特征提取、计算特征权重[17]等步骤后分别得到它们的向量表示,本文利用向量夹角的余弦计算得到文献之间的相似度[16],然后以 K-Means 方法进行文本的聚类。聚类方法的详细描述与性能测评可参见文献[17]。

(3)聚类结果描述

如上所述,传统聚类算法直接用于文本聚类上,存在的突出问题是算法的有效性问题,因为传统的聚类算法只对待聚类对象进行聚类,不负责对聚类后生成的类簇进行概念描述和语义解释。因此,必须针对

文本聚类的特别要求,探寻专门解决文本聚类描述这一问题的方法。自动化的聚类描述主要从聚类生成的类簇中提取重要的词语,根据聚类算法的不同,相应的词语重要性计算方法也有所不同[18]。

本文采用一种增强聚类结果的可理解性与可读性的算法,即基于支持向量机(SVM)[19]的文本聚类结果描述算法。该方法所使用的候选描述词特征有:DF * ICF,描述词的文档频率与逆类簇频率;TF * IDF(t_{ij}),描述词 t_{ij} 在 $\vec{C}_i$ 内的 TF * IDF(聚类前)均值;DEP,描述词在当前类簇内部首次出现位置的均值;Position_global,描述词全局位置特征,包括当前类簇内部描述词在题名(Title)、摘要(Abstract)、章节标题(Heading)、文章第一段(FirstPara)、文章最后一段(LastPara)位置出现的文档比率;POS,描述词词性;LEN,描述词长度。实验结果表明,基于支持向量机的聚类描述算法所取得的效果要优于常规的聚类结果描述方法,准确率约为62%[20]。该部分详细算法描述可参见文献[20]。

本文通过机器学习获得聚类结果的描述,并通过逐层聚类生成聚类描述的类别体系[20](本文生成的分类体系的层次为两层)。在此基础上对分类体系进行人工审定,生成供用户导航的分类层次结构。例如,人工审定后的CNKI"房地产"主题数字图书馆就是两层聚类浏览层次结构的浏览界面,粗体标题部分为一级类目(括号内数字表示该类目下的文档数目),每个类目下的标题表示该类目的子类目。

3.3　主题数字图书馆的实现结果与评估

(1)主题数字图书馆的实现结果

本文按图1所示的主题数字图书馆设计框架图,以CNKI期刊数据库资源为基础,通过主题采集获得主题库。主题库经主题提取得到能表达每篇文献主题的关键词集合,以此为基础进行文本聚类,生成类簇描述,然后对类簇描述进行人工审定,生成每一主题库的一级聚类浏览层次结构。本文对子类目下的文献再次聚类、聚类描述及类目的人工审定,最终生成两层主题导航结构。图2给出的房地产类主题库(包括160,407篇相关文献)的主题导航与查询结果样例中,用户在进行房地产主题数字图书馆的浏览或查询时,可根据左边聚类导航条在某一子主题下,进行该子主题的聚类浏览和检索。

本文已经开发完成"足球知识"(包括21,422篇文献)、"煤炭知识"(包括141,004篇文献)、"房地产知识"等十余个CNKI主题数字图书馆。目前上线运行的主题数字图书馆有四大名著专题和足球专题(登陆网址为:http://topic.cnki.net)。系统上线运行近两年,日平均用户访问量可观。

(2)主题数字图书馆性能评估

由于评价主题数字图书馆的信息检索和服务性能的标准不确定,难以从定量角度进行评估,本文主要从聚类浏览结果的角度来评估基于主题聚类的主题数字图书馆系统的性能,即:对生成的两层聚类浏览层次结构进行评估。需要指出的是,聚类浏览层次结构评估结合了类目结构层次的宏观评价与类目描述质量的微观评价,因此评估难度比仅仅评估聚类描述要大得多。为此,本文针对主题数字图书馆的聚类浏览结构,设计了一个评估问题域,包括3个问题(见表1)。本文通过5名志愿者对这三个问题的打分情况进行评估,表1给出这三个评估标准的打分规则。其中,聚类浏览结构的均衡度是指聚类浏览结构中,每个类目下聚类样本分布数目的均衡程度;相关度是指聚类描述与当前主题数字图书馆的主题相关程度;聚类浏览结构的总体效果是要求每个志愿者对聚类浏览结构作一个总体评价。

表1　聚类浏览结构评估方法

问题号	评估标准	打分规则
1	聚类浏览结构的均衡度	均衡(2分),较均衡(1分),不均衡(0分)
2	类目描述的相关度	相关(2分),较相关(1分),不相关(0分)
3	聚类浏览结构的总体效果	好(7~10分),一般(4~7分),不好(0~4分)

为了更好地考察主题聚类方法的效果,本文引入一个基准方法(BaseLine),进行聚类浏览层次效果的对比分析。基准方法的思路为:直接统计文档集合的关键词,取出现频次最大的前N个关键词作为一级类目,然后在每个类目下再次进行关键词词频统计(当前一级类目关键词不再参加统计),将出现频次最大的前M个关键词作为一级类目下的二级类目。

表 2 聚类浏览结构评估结果

项目 / 类别	均衡度		相关度		总体效果评估	
	TC	BL	TC	BL	TC	BL
房地产	1.57	1.12	1.78/1.82	1.48/1.50	8.12	6.92
煤炭	1.68	1.21	1.72/1.78	1.68/1.71	8.20	7.04
足球	1.45	1.01	1.62/1.67	1.59/1.62	7.38	5.94
航空航天	1.64	0.92	1.53/1.61	1.37/1.45	7.82	5.46
汽车	1.49	0.94	1.61/1.70	1.52/1.58	7.49	5.59
平均值	1.57	1.04	1.65/1.72	1.53/1.57	7.80	6.19

本文安排 5 名志愿者根据表 1 给出的评估方法,对“房地产知识”、“煤炭知识”、“足球知识”、“航空航天知识”、“汽车知识”共 5 个主题数字图书馆进行聚类浏览层次结构的评估。统计他们的打分结果(见表 2),其中 TC、BL 分别代表所评价的对象,即基于主题聚类方法的浏览层次结构和基于基准方法的浏览层次结构,志愿者评价时不知道具体评价对象名称。相关度的评估细分为一级类目描述的相关度评估与二级类目描述的相关度评估,如“房地产知识”主题数字图书馆统计结果为 1.78/1.82,代表一级类目总体相关度为 1.78,二级类目为 1.82(见表 2)。

从表 2 看出,以均衡度为角度,基于主题聚类方法生成的聚类浏览层次结构优于基准方法生成的层次结构,前者的均衡度达 1.57,而后者为1.04。两者的结果都比较均衡,但后者直接以关键词词频统计生成聚类层次结构,没有很好地解决交叉类目情况,造成有些同级类别的重复。例如,“房地产知识”主题数字图书馆中,基准方法存在“房地产商”与“房地产开发商”这样语义非常接近的一级类目描述词。主题聚类方法以主题提取为基础,进行主题聚类和聚类描述,能够克服这一问题。因此,主题聚类方法生成的聚类浏览层次结果相对比较均衡,适合用户对某一专题的聚类浏览和检索。

主题聚类方法相关度评价结果为:一级类目相关度为 1.65,二级类目相关度为 1.72,该结果均优于基准方法,但两者差异不及均衡度明显。由于主题聚类方法中使用候选描述词的多个特征,利用支持向量机相对高效地提取出类簇描述词,而基准方法仅以候选描述词出现频次作为特征,无法避免交叉类目、类目描述同义或近义等情况,因此主题聚类方法生成的类别描述,相关度要优于基准方法。

从表 2 还看出,无论是主题聚类方法还是基准方法,二级类目的相关度评估结果均优于一级类目的结果。主题库经过第一层聚类生成的每个类簇中的不相关文献相对减少,若对每一个类簇下的文献再次聚类(即第二层聚类),生成二级聚类类目与其对应的一级类目较相关,因此二级类目相关度评估值略高于一级类目相关度评估值。

从表 2 可知,志愿者对主题聚类方法生成聚类浏览层次结构的总体评估效果为 7.80(7 ~ 10 分之间),对基准方法的总体效果评估为 6.19(4 ~ 7 分之间),即:用户认为主题聚类方法的总体效果较优,而认为基准方法的效果一般。

综上,以聚类浏览层次结构为视角,从均衡度、相关度、总体效果评估来看,主题聚类方法均优于基准方法。同时从评估结果也可看出,主题聚类方法并未达到最理想的结果。本文的下一步工作包括进一步提高聚类浏览层次均衡度、相关度,设计更加合理的评估方法等。

4 结束语

主题数字图书馆是一种基于主题特征的专业领域数字图书馆。本文提出基于主题聚类的信息资源组织模式,并将该模式用于主题数字图书馆的构建。首先从大规模的科技文献集合中,利用规则与统计结合的文本自动分类方法,获取特定领域的科技文献子集。然后利用机器学习方法抽取每篇文献的关键词,以此为基础进行主题聚类,自动生成特定领域的分类体系。最后,分类体系经人工审定后,生成供用户导航的分类层次结构,结合全文检索从而实现

基于主题聚类的主题数字图书馆系统。主题聚类与一般文本聚类不同,它是聚类样本在主题提取的基础上进行的聚类,二者在理论方法和实际性能上的差别,还有待进一步深入研究和分析。本文下一步工作还包括:研究主题聚类中各步骤之间的相互影响关系,寻求主题聚类全局最优化的方法;针对大规模样本集,包括 Web 文本资源,进行主题聚类方法的测试;研究基于大规模数据集的主题数字图书馆的构建问题,并对主题数字图书馆的应用效果评估进行深入研究等。

参考文献:

[1] 马张华, 陈文广,金海燕,等. 基于控制词集的中文信息动态自动聚类研究[J]. 大学图书馆学报, 2006, 24(6): 54 - 60.

[2] Kang S S. Keyword-based Document Clustering [C]. Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages, Sapporo, Japan, 2003: 132 - 137.

[3] Chang H - C, Hsu C - C. Using Topic Keyword Clusters for Automatic Document Clustering [J]. IEEE Transactions on Information and Systems, 2005, E88 - D: 1852 - 1860.

[4] 孙学刚, 陈群秀, 马亮. 基于主题的 Web 文档聚类研究[J]. 中文信息学报, 2003, 17(3): 21 - 26.

[5] Zhao Y, Karypis G. Topic-driven Clustering for Document Datasets [C]. Proceedings of the Fifth SIAM International Conference on Data Mining, St. Louis, Missouri, 2005: 358 - 369.

[6] 赵世奇, 刘挺, 李生. 一种基于主题的文本聚类方法[J]. 中文信息学报, 2007, 21(2): 58 - 62.

[7] Andreas R., Dieter M. SOMLib: A Digital Library System Based on Neural Networks [C]. Proceedings of the Fourth ACM conference on Digital Libraries, Berkeley, CA, USA, 1999: 240 - 241.

[8] Cutting, D. R., Karger, D. R, Pedersen, J. O. and Tukey, J. W. Scatter/Gather: A cluster-based approach to browsing large document collections [C]. Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92), Copenhagen, Denmark, 1992: 318 - 329.

[9] 中国学术期刊(光盘版)电子杂志社. 中国知网[OL]. [2006-07-10]. http://www.cnki.net.

[10] Yang Y. An Evaluation of Statistical Approaches to Text Categorization [J]. Journal of Information Retrieval, 1999, 1(1 - 2): 69 - 90.

[11] Salton G, Yang C S, Yu C T. A Theory of Term Importance in Automatic Text Analysis [J]. Journal of the American society for Information Science, 1975, 26(1): 33 - 44.

[12] Hulth A. Improved Automatic Keyword Extraction Given More Linguistic Knowledge [C]. Proceedings of the 2003 Conference on Emprical Methods in Natural Language Processing, Sapporo, Japan, 2003: 216 - 223.

[13] Turney P D. Learning to Extract Keyphrases from Text. NRC Technical Report ERB - 1057[R]. National Research Council, Canada. 1999: 1 - 43.

[14] 韩客松, 王永成. 中文全文标引的主题词标引和主题概念标引方法[J]. 情报学报, 2001, 20(2): 212 - 216.

[15] 张庆国, 薛德军, 张振海, 张君玉. 海量数据集上基于特征组合的关键词自动抽取[J]. 情报学报, 2006, 25(5): 587 - 593.

[16] 章成志, 师庆辉, 薛德军. 基于样本加权的文本聚类算法研究[J]. 情报学报, 2008, 27(1): 42 - 48.

[17] G. Salon and M. J. Mcgill. Introduction to Modern Information Retrieval[M]. McGraw-Hill Book Co., New-York, 1983.

[18] Tseng Y - H, Lin C - J, Chen H H, Lin Y - H. Toward Generic Title Generation for Clustered Documents [C]. Proceedings of the 3rd Asia Information Retrieval Symposium, Singapore, 2006: 145 - 157.

[19] Vapnik V. The Nature of Statistical Learning Theory [M]. New York: Springer-Verlag, 1995: 1 - 175.

[20] 章成志. 主题聚类及其应用研究[D]. 南京:南京大学, 2007.

章成志　南京理工大学信息管理系讲师,博士,中国科技信息研究所博士后。通讯地址:南京。邮编 210094。

张庆国　中国学术期刊(光盘版)电子杂志社,硕士。通讯地址:北京。邮编 100084。

师庆辉　中国学术期刊(光盘版)电子杂志社,学士。通讯地址同上。

(收稿日期:2008-04-22)