中 文 信 息 学 报
第 20卷 第 3期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 20 No. 3

[综述]文章编号: 1003 - 0077(2006)03 - 0055 - 08

文档聚类综述*

刘远超, 王晓龙, 徐志明, 关 毅

(哈尔滨工业大学 计算机科学与技术学院, 黑龙江 哈尔滨 150001)

摘要: 聚类作为一种自动化程度较高的无监督机器学习方法, 近年来在信息检索、多文档自动文摘等领域获得了广泛的应用。本文首先讨论了文档聚类的应用背景和体系结构, 然后对文档聚类算法、聚类空间的构造和降维方法、文档聚类中的语义问题进行了综述。最后还介绍了聚类质量评测问题。

关键词: 计算机应用; 中文信息处理; 综述; 文档聚类; 降维; 概念相关; 聚类算法

中图分类号: TP391 **文献标识码**: A

A Survey of Document Clustering

LIU Yuan-chao, WANG Xiao-long, XU Zhi-ming, GUAN Yi

(School of Computer Science and Technology, Haerbin Institute of Technology, Haerbin 150001)

Abstract: As an unsupervised machine learning method, document clustering has been widely used in many NLP applications such as information retrieval, automatic multi-document summarization and etc. In this paper the background and the architecture of document clustering is discussed firstly, and then some related problems are surveyed which includes clustering algorithm, feature space construction, dimension reduction and the semantic problem. In the end this paper introduces the evaluation of cluster quality.

Key words: computer application; Chinese information processing; overview; document clustering; dimension reduction; concept relevance; clustering algorithm

1 文档聚类的应用背景和体系结构

作为一种无监督的机器学习方法, 聚类技术已经成为对文本信息进行有效地组织、摘要和导航的重要手段, 为越来越多的研究人员所关注[1~3]。文档聚类的主要应用点包括: ①文档聚类可以作为多文档自动文摘等自然语言处理应用的预处理步骤。其中比较典型的例子是哥伦比亚大学开发的多文档自动文摘系统 Newsblaster[4]。Newsblaster将每天发生的重要新闻进行聚类处理, 并对同主题文档进行冗余消除、信息融合、文本生成等处理, 从而生成一篇简明扼要的摘要文档; ②对搜索引擎返回的结果进行聚类, 使用户迅速定位到所需要的信息[5]。比较典型的系统有 vivisimo (http://www.vivisimo.com)和 infonetware (http://www.infonetware.com)等。系统允许用户输入检索关键词, 而后对检索到的文档进行聚类处理, 并输出各个不同类别的简要描述, 从而可以缩小检索的范围, 用户只需关注比较有希望的主题。另外这种方法也可以为用户二次检索提供线索; ③对用户感兴趣的文档 (如用户浏览器 cache中的网页)

* 收稿日期: 2005 - 07 - 01 定稿日期: 2005 - 10 - 08
基金资助: 国家自然科学基金重点资助项目 (60435020)
作者简介: 刘远超 (1971—), 男, 讲师, 博士生, 主要研究方向为自然语言理解、文本挖掘.

聚类,从而发现用户的兴趣模式并用于信息过滤和信息主动推荐等服务[6];④聚类技术还可以用来改善文本分类的结果,如俄亥俄州立大学的 Y. C. Fang等人的工作[7];⑤数字图书馆服务。通过 SOM(自组织映射)等方法,可以将高维空间的文档向量拓扑保序地映射到二维空间,使得聚类结果可视化和便于理解,如 SOMlib[8]系统;⑥文档集合的自动整理。如 Scatter/Gather[9]是一个基于聚类的文档浏览系统。而微软的 Ji-Rong Wen[10]等人则利用聚类技术对用户提出的查询记录进行聚类,并利用结果更新搜索引擎网站的 FAQ。

文档聚类又分为硬聚类和软聚类。前者每个文档只能属于一类,即:

$$C_1 \cup C_2 \cup \cdots \cup C_k = DC, C_i \cap C_j = \Phi, \text{其中}, 1 \leq i \neq j \leq k \quad (1)$$

而如果是软聚类,则文档集合被划分为 k个可能相交的文档子集,即每个文档可能属于多个类别。

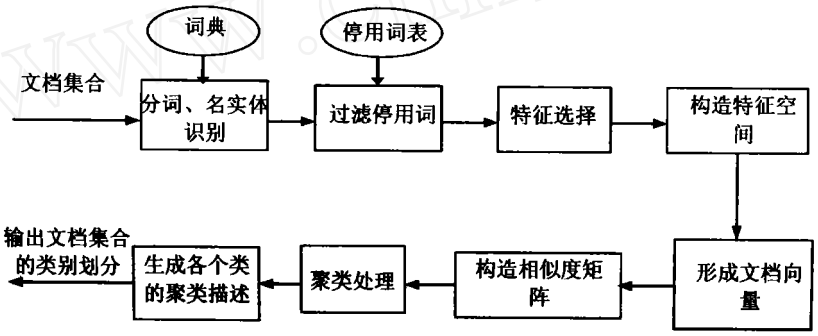

图 1 文档聚类的一般体系结构

文档聚类算法的输出一般为文档集合的一个划分。这种划分的形式也有可能是一个层次结构(如 AHC算法)或者二维平面图(如 SOM神经网络)。文档聚类一般采用图 1所示的体系结构。系统首先需要构造聚类特征空间,并将所有文档表示为特征空间上的向量。由于聚类迭代过程中经常需要根据文档(或者中间类)之间的相似度来进行合并或者划分等操作,因此为提高运行效率,可以预先生成文档之间的相似度矩阵。系统运行过程中,可以到矩阵中检索任何两个文档的相似度。

聚类描述是帮助用户迅速确认生成的文档类相关与否的重要信息。Hearst和 Pedersen等人[11]利用聚类关键词来形成类别描述。Anton V. Leouski和 W. Bruce Croft等人[12]则提出用关键词短语(名词短语)来进行类别描述,因为关键词短语比词表达的信息更加丰富。他们主要根据词的文档频率来评价词的重要性。华盛顿大学的 Oren Zamir和 Oren Etzioni等人开发的 Grouper[13]也从文档集合中抽取关键词短语来作为类别描述。新加坡国立大学的 Dell Zhang[14]提出一种基于语义的层次在线信息检索聚类系统 SHOC,对 O. Zamir和 O. Etzioni的工作进行了扩展。该系统实现了关键词短语的自动抽取和基于概念的正交聚类,而且可以同时支持中文和英文两种语言。

作为一种自然语言处理应用,文档聚类具有高维和与语义相关的特点,因此影响文档聚类结果的因素除了文档聚类算法的选择以外,还包括语义问题的处理和降维问题。这些问题也是目前文档聚类研究中的难点和热点。本文将在下面围绕这几方面展开综述。

2 文档聚类算法

如前所述,文档聚类有多种应用,不同的应用对聚类质量、效率以及结果可视化程度等方面往往都有特定的要求,因此要根据应用场合和目的选用适合的聚类算法。表 1对一些常用

的文档聚类算法进行了分析和比较。

AHC(层次聚合聚类)算法[15]首先假设所有文档自成一类,然后将最相似的两类合并,并继续这一过程,直到最后将所有文档合并为一类,因而可以形成一棵聚类树。文献[16]认为在信息检索的应用背景下,层次聚合聚类在检索相关文档方面要比基于划分的算法要好。一般说来,聚合过程中会有一次的聚类结果比较符合真实的类别划分,这种最佳划分结果可以根据阈值或者聚类熵确定。Jung在他的博士论文中指出通过评价划分的聚类熵[17]有助于解决这一问题,聚类熵定义为:

$$En = \left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} e(p_i^{(j)}, p_0^{(j)})\right) + \sum_{j=1}^{k} e(p_0^{(j)}, c_0) \qquad (2)$$

其中公式右边的第一项代表类内熵值,第二项代表类间熵值。公式 c_0 中表示所有样本的中心,$p_i^{(j)}$ 表示第 j类的第 i个样本,$p_0^{(j)}$ 表示第 j类样本的中心,k值为聚类的个数。可以取 En 最小时的层次划分作为层次聚类的结果。

层次聚合聚类方法的计算复杂度一般为 $o(n^2)$,其中 n为输入文档的个数。

随着网络信息量的爆炸增长,要求聚类算法效率要高,且效果要好。因此也相应出现了一些高效方法来处理聚类问题,其中包括基于划分的方法、自组织映射网络等。

k-means[18]是一种典型的基于划分的方法。其基本原理是首先选择 k个文档作为初始的聚类点,然后根据簇中对象的平均值,将每个文档(重新)赋给最类似的簇,并更新簇的平均值,然后重复这一过程,直到簇的划分不再发生变化[19]。k-means的算法复杂度为 $O(klN)$,其中 l为迭代次数,N为文档个数,k为类别个数。

k-means算法本质上是一种贪心算法。可以保证局部最小,但是很难保证全局最小。另外该方法需要预先指定 k值和初始划分,从而容易使聚类结果受到影响。为此人们提出了一些相应的解决方法。如可以将算法执行多次,取最好的一次作为最终结果。文献[20]采用遗传算法来优化 k值。Tao Li等人提出一种聚类算法 ASI[21],对如何自动预测聚类的个数进行了研究。在已知 k值的情况下,如何获得初始聚点也是一个值得关注的问题。Likas et al 等人提出的 global k-means方法[22]对初始聚点的选择提出了新的方案,而不是随机选择聚点。比较有效的方法还有基于最小最大原则的方法和基于密度的方法等[23]。

另外一种值得关注的文档聚类方法是基于 SOM 神经网络的方法,该方法由 T. Kohonen[24]首先提出。使用 SOM 方法进行文档聚类的基本原理是:

(1)首先对输出层各个神经元所代表的权值向量赋小的随机数,并归一化处理。神经元的向量维数与输入文档向量的维数相同;

(2)从训练集中随机选取文档向量,作为 SOM 网络的输入;

(3)计算输入文档向量与各神经元向量的相似度,相似度最大的神经元将获胜;

(4)获胜神经元及其邻域内的神经元调整权值,权值调整的幅度一般采用随时间单调下降的退火函数。

通过调整权值,获胜者及其邻域内的神经元和输入文档模式更加接近,因此使这些神经元以后对相似输入模式的响应得以增强。通过使用大量训练文本训练网络,最后使输出层各节点成为对特定模式类敏感的神经细胞,对应的向量成为各个输入模式类的中心向量。

SOM 输出节点的个数与输入文档集合的类别个数有关系。SOM 聚类的难点是如何设置输出层的节点个数。如果节点数少于类别数,则不足以区分全部模式类,结果将使相近的模式类合并为一类。如果节点数多于类别数,则将导致类别划分过细,从而对聚类质量和网络的收

效效率产生影响。

表 1 几种主要的文档聚类方法

聚类算法	基本原理	代表性的方法
基于划分的方法 (Partition-based clustering)	首先得到初始的 k个划分,然后通过迭代,将文档从一个中间类转移到另一个类中,以改进聚类的质量。	k-means、k中心点、CLARANS
层次聚类 (Hierarchy-based clustering)	对文档集合进行层次分解。可以分为自底向上和自上向下两种类型。	AHC、CURE, Chameleon, BIRCH
基于模型的方法 (Model-based clustering)	从文档集合中学习一个模型,每个模型代表一个文档类。并优化给定的数据和数学模型之间的适应性。	Bernoulli model, vMF, Gaussian
SOM 神经网络 (Self organizing map)	通过对网络进行训练,将文档从高维空间向二维空间进行保序映射。	SOM lib, GH-SOM, WEBSOM
基于蚁群的方法 (Ants-based clustering)	在二维空间内随机放置文档对象,蚁群根据文档与其邻域文档的相似信息,可以拾起文档并在适当的位置放下文档。	CSI, ACCL, AntClust

SOM 可以将高维空间的数据转化为二维空间,并且输入文档彼此之间的相似性在二维离散空间得到很好的保持。另外该方法还具有对噪声不敏感和聚类质量较高的特点。标准 SOM 的算法复杂度为 $O(k'mn)$,其中 k'为神经元个数,m 为训练次数 (需随机输入 $m \cdot n$次样本进行训练)。网络输出层的神经元个数 k'一般大于可能的类别个数,m 一般也大于 k-means 算法的迭代次数,因此 SOM 的运行时间往往略高于 k-means算法。但二者基本上都具有线性的复杂度。研究表明,SOM 在输出层神经元得到充分利用时聚类质量明显优于 k-means,使得这种方法成为一种值得关注和进一步研究的文档聚类方法。

目前基于 SOM 的文档聚类方法在数字图书馆等领域得到了较好的应用。传统 SOM 的网络结构、神经元的数目和分布在训练之前需要预先确定,因而难以做到对输入数据的自适应性,并且输入数据的层次关系也无法体现。层次 SOM 特征图则克服了这一问题。其关键思想由 R. Miikkulainen于 1990年提出,可以表示输入数据的层次关系。但是该模型中构成层次图的不同 SOM 的大小和层次的深度必须预先指定。因此为了获得理想的结果,必须了解输入数据的结构,这影响了聚类的无指导特性。文献 [25]提出的 GH-SOM 模型对层次 SOM 进行了改善。在 GH-SOM 中,层次的深度和每一层的神经元数目都可以自适应的调整,使得灵活性大大增强。将 SOM 应用在文档聚类的文献还有 [26, 27]等。

上面介绍的是目前文档聚类中的几种主流方法。文献 [28]对不同的聚类算法及其应用进行了对比研究。

3 文档聚类中的概念相关问题

和很多自然语言处理问题一样,文档聚类也是和语义密切相关的。然而传统上大多数文档聚类方法都单纯利用词频信息构造相似度矩阵。在这种模式下,两篇文档相似度较大的一个主要原因是它们之间存在较多的公共词。因此即使人们认为应该属于同类的文档可能由于相似度较小而被误判为不同类。在一些实际应用中,人们要求聚类算法能深入到概念一级。很多文档聚类算法的研究评测采用的语料中同类文档的主题分布一般比较广泛,且公共词较少。如英文文档聚类常用的 *Reuters* 21578、20*Newsgroups*,中文文档聚类采用的北大分类语料等。采用单纯词频向量进行聚类,由于同类文档之间相似度较小,将导致聚类结果较差。

由于语言表达形式的多样性,即使同一概念,往往有多种不同的表达方式,如“天气”和

“气候”二词在语义上是很接近的。另外,词与词之间往往存在较强的语义关联关系,如各种上下位关系、同义词、反义词等等。这些复杂的关系往往是人工标注同类语料的重要依据。因此,单单依靠特征词的重复而产生的频率信息是完全不够的。如果将特征项映射到概念级,将有助于加强同一类别文档的聚合能力[29]。德国卡尔斯鲁厄大学的 Andreas Hotho和 Steffen Staab等人提出一种基于本体的文档聚类算法[30]。其基本思想是将词映射为概念,如具体的旅店名称统统映射为“旅店”一词。他们主要是利用 wordnet作为知识源来解决语义问题[31]。美国 NEC国家实验室的 Wei Xu[32]等人则提出一种概念分解的聚类方法。通过概念映射或者文档的概念词扩展处理,将有助于使同类文档之间的相似度加大,而缩小不同类文档之间的相似度。这符合聚类假设,也和人们的期望一致。笔者利用知网作为知识源,对文档中的高频词进行相关词扩充,并据此构造文档的概念词矢量用于聚类,取得了较好的效果,表 2和表 3是相关词扩充前后的类别相似度对比(采用中文分类语料)。可以看出,经过相关词扩充处理,使得同类文档之间的相似度明显加大,因而可以更好的满足聚类假设,有利于改善聚类结果。

表 2 类别相似度的取值对比情况(概念扩充前)

	C1	C2	C3	C4	C5
C1	0.1316	—	—	—	—
C2	0.1266	0.1643	—	—	—
C3	0.1044	0.1323	0.1507	—	—
C4	0.0167	0.1235	0.0839	0.1531	—
C5	0.0839	0.1061	0.0930	0.0962	0.1478

表 3 类别相似度的取值对比情况(概念扩充后)

	C1	C2	C3	C4	C5
C1	0.3867	—	—	—	—
C2	0.2389	0.3488	—	—	—
C3	0.1008	0.2102	0.7311	—	—
C4	0.1305	0.1457	0.0254	0.4173	—
C5	0.1526	0.0823	0.0724	0.0325	0.6250

4 特征空间的降维处理

由于文档聚类的特征空间维数较高,以及聚类算法采用多次迭代求精的策略,使得文档聚类算法的时间开销相对较大。特别是当文档聚类用于信息检索等应用时,其实时性的提高就更是一个重要的课题。为了提高文档聚类系统的运行效率,一般有两种途经:一是采用合适的降维方法;二是采用高效的聚类算法。通过选用 k-means等算法,可以改善聚类算法自身的效率。而文档聚类中的降维问题近年来也引起了研究人员的广泛关注。通过对聚类空间进行有效降维,将在不影响聚类质量的前提下,大大改善聚类的效率。因而研究合理的降维方法对于文档聚类的实际应用是非常必要的。

Mark P. Sinka和 David W. Corne等人对英文文档的聚类结果表明,对文档进行停用词过滤将会提高聚类的效果,而词形还原处理则对聚类的效果有微弱的负面影响[33]。另外他们还指出,在获取集合中所有的唯一词集合以后,构造文档的特征空间时,可以只保留其中频率较高的部分词。他们的实验表明,与人们的常规认识相反,并不是将所有的唯一词都作为特征空间,其聚类效果就最好。当选择 1%的高频词时,聚类效果反而最好。这是因为如果保留比较

多的特征词,可能将具有较强类别辨识能力的词纳入特征空间,也有可能会将辨识能力不够强的词(如某些类停用词)吸收进来。Anton V. Leouski和W. Bruce Croft等人指出每篇文档只保留50～100个词可以基本满足聚类的需要,而不会对聚类的结果发生影响。他们还指出,单纯使用文档频率来进行文档表示已经足够,而不需要更加复杂的表示方法。文献[34]提出将文档中抽取出的关键词作为特征。文献[35]提出采用潜在语义索引(LSI)方法压缩聚类特征空间的维数。Kristina Lerman的实验表明,利用LSI方法可以取得比使用相同聚类算法但在没有经过压缩的空间进行聚类更好的聚类质量,并且在压缩后的特征空间进行聚类运算所节省的时间弥补了由于获取压缩特征空间所花费的计算开销。文献[36]对特征词的重要性进行了研究,指出特征词的选取对聚类质量有较大影响。Liu[37]等人提出一种基于词同现频率的特征选择方法。除此之外,ZhengYu Niu[38]和Stanislaw Osiński[39]等也对特征选择进行了研究。特征选择对概念映射理论下进行面向语义的文档聚类也是必要的,为了提高效率,可以只选择原文中的部分高频词进行概念扩充,另外还有必要对扩充后形成的向量作进一步压缩。

5 聚类质量评测

为了评价聚类系统的整体性能,可以采用两种常用的指标:纯度[40]和F值[41]。采用的数据一般是人工分好类的文档集合,如各种常用的文本分类语料以及包含多个不同主题的web文档集合。

对于聚类后形成的任意类别 r,聚类的纯度定义为

$$P(S_r) = \frac{1}{n_r}\max(n_r^i) \tag{3}$$

整个聚类结果的纯度定义为

$$Purity = \sum_{r=1}^{k}\frac{n_r}{n}P(S_r) \tag{4}$$

这里,n_r^i是属于预定义类 i且被分配到第 r个聚类的文档个数,n_r为第 r个聚类类别中的文档个数。

而 F值的定义则参照信息检索的评测方法,将每个聚类结果看作是查询的结果,这样,对于最终的某一个聚类类别 r和原来的预定类别 i,

$$recall(i, r) = n(i, r)/n_i \tag{5}$$

$$precision(i, r) = n(i, r)/n_r \tag{6}$$

这里 $n(i, r)$是聚类 r中包含类别 i中的文档的个数,n_r是聚类形成的类别个数,n_i是预定义类别的个数。则聚类 r和类别 i之间的 f值计算如下:

$$f(i, r) = (2 \cdot recall(i, r) \cdot precision(i, r))/(precision(i, r) + recall(i, r)) \tag{7}$$

最终聚类结果的评价函数为

$$F = \sum_i \frac{n_i}{n}\max\{f(i, r)\} \tag{8}$$

这里 n是所有测试文档的个数。值得指出的是,通过以上这两种方法获得的聚类评价只是对数据集作一次划分的评价。为了客观评价聚类算法的性能,有必要进行多次聚类获得其评价结果,并用其均值(方差)来评价算法。

6 结论

本文对文档聚类中比较重要的几个问题:文档聚类算法、聚类空间的构造和降维方法、文档聚类中的语义问题进行了综述,同时也介绍了聚类质量评价问题。随着网络信息的飞速增长,文档聚类这样的研究显得越来越重要了。希望通过本文的论述,能为相关研究起到抛砖引玉的作用。

参 考 文 献:

[1] 马帅,王腾蛟,等. 一种基于参考点和密度的快速聚类算法[J]. 软件学报. 2003, 14(6): 1089 - 1095.

[2] 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.

[3] 吴斌,傅伟鹏,史忠植,等. 一种基于群体智能的web文档聚类算法[J]. 计算机研究与发展, 2002, 39(11): 1429 - 1435.

[4] Regina Barzilay, Min-Yen Kan, and Kathleen R. McKeown. Simfinder: A Flexible Clustering Tool for Summarization[A]. In proceedings of the Workshop on Summarization in NAACL 01[C]. Pittsburg, Pennsylvania, USA: June 2001.

[5] Zheng Chen, Wei-Ying Ma, Jinwen Ma. Learning to Cluster Web Search Results[A]. In: proceedings of the 27th Annual International ACM SIGIR Conference[C]. Sheffield, South Yorkshire, UK, July 2004, 210 - 217.

[6] 林鸿飞,马雅彬. 基于聚类的文本过滤模型[J]. 大连理工大学学报. 2003, 42(2).

[7] Y. C. Fang, S. Parthasarathy, F. Schwartz. Using Clustering to Boost Text Classification[J]. In: proceedings of the IEEE ICDM Workshop on Text Mining, Maebashi City, Japan, 2002.

[8] A. Rauber, and M. Frühwirth. Automatically Analyzing and Organizing Music Archives[A]. In: proceedings of the 5. European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001)[C]. Darmstadt, Germany, 2001.

[9] Cutting, D., Karger, D., and etc. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections[A]. SIGIR '92, 1992[C]. 318 / 329.

[10] JR Wen, JY Nie, HJ Zhang. Clustering User Queries of a Search Engine[A]. The Tenth International World Wide Web Conference[C]. Hong Kong May 1 - 5, 2001.

[11] Anton Leuski and James Allan. Improving Interactive Retrieval by Combining Ranked Lists and Clustering[A]. In: proceedings of RIAO 2000[C]. Paris, France, April 12 - 14, 2000, 665 - 681.

[12] Anton V. Leouski and W. Bruce Croft. An Evaluation of Techniques for Clustering Search Results[A]. Technical Report IR - 76, Department of Computer Science, University of Massachusetts, Amherst, 1996.

[13] Http://www.cs.washington.edu/research/clustering

[14] Dell Zhang. Semantic, Hierarchical, Online Clustering of Web Search Results[A]. In: proceedings of the 6th Asia Pacific Web Conference (APWEB)[C]. Hangzhou,. China, April 2004.

[15] P. H. Sneath and R. R. Sokal. Numerical Taxonomy[M]. Freeman, London, UK, 1973.

[16] P. Willett. Recent trends in hierarchic document clustering: a critical review[J]. In: Information Processing and Management, 24(5): 577 - 597, 1988.

[17] Yunjae jung. Design and Evaluation of Clustering Criterion for Optimal Hierarchical Agglomerative Clustering[D]. Phd. thesis. University of Minnesota. 2001.

[18] 行小帅,潘进,焦李成. 基于免疫规划的K-means聚类算法[J]. 计算机学报, 2003, 26(5): 605 - 610.

[19] 陈浩,何婷婷,姬东鸿. 基于k-means聚类的无导词义消歧[J]. 中文信息学报, 2005, 19(4): 10 - 16.

[20] A. Casillas, M. T. González de Lena and R. Martínez. Document clustering into an unknown number of clusters using a Genetic Algorithm[A]. International Conference on Text Speech and Dialogue TSD, 2003.

[21] Tao Li. Document clustering via Adaptive Subspace Iteration [A]. In: proceedings of the 12th ACM International Conference on Multimedia[C]. New York, USA, 364 - 367, 2004.

[22] A. Likas, N. Vlassis, and J. J. Verbeek. The global k-means algorithm. Pattern Recognition [J]. Vol 36, 2003, 451 ~ 461.

[23] 范金城,梅长林. 数据分析 [M]. 科学出版社. 2002年 7月第一版.

[24] T. Kohonen. Self-organized formation of topologically correct feature maps[J]. Biological Cybernetics, 43: 59-69, 1982.

[25] Michael Dittenbach, Dieter Merkl, Andreas Rauber. The Growing Hierarchical Self Organizing map [A]. In: proceedings of the Int'l Joint Conference on Neural Networks (IJCNN'2000) [C]. Como, Italy, July 24-27, 2000.

[26] X. Lin, D. Soergel, and G. Marchionini. A self-organizing semantic map for information retrieval[A]. In: proc. ACM SIGIR int'l conf in information retrieval (SIGIR'91) [C]. Chicago, Illinois, 1991.

[27] K. Lagus, T. Honkela, S. Kaski, and T. Kohonen. Self-organizing maps of document collections: A new approach to interactive exploration[A]. In: proc int'l conf knowledge discovery and data mining (KDD'96) [C]. Portland, Oregon, 1996.

[28] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review[J]. ACM Computing Surveys, 31 (3): 264 ~ 323, 1999.

[29] 史忠植. 知识发现 [M]. 清华大学出版社. 2002年 1月第一版.

[30] A. Hotho, A. Maedche, S. Staab. Ontology-based Text Clustering [A]. IJCAI-2001 Workshop.

[31] Andreas Hotho. Wordnet improves Text Document Clustering [A]. In: proc. of the SIGIR 2003 Semantic Web Workshop [C]. Toronto, Canada, 2003.

[32] Wei Xu, Yihong Gong. Document Clustering by Concept Factorization [A]. In proceedings of the 27th ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Sheffield, UK, 2004.

[33] Mark Sinka and David Corne. A Large Benchmark Dataset for Web Document Clustering[J]. In Soft Computing Systems: Design, Management and Applications, Vol 87 of Frontiers in Artificial Intelligence and Applications, pages 881 ~ 890, 2002.

[34] Seung-Shik Kang. Keyword-based Document Clustering [A]. The 6th International Workshop on Information Retrieval with Asian Languages[C]. IRAL2003, p132-137, July, 2003.

[35] Kristina Lerman. Document Clustering in Reduced Dimension Vector Space [A]. In: proceedings of CSAW'[C]. 04, 2004.

[36] Christian Borgelt and Andreas Nürnberger. Experiments in Document Clustering using Cluster Specific Term Weights [A]. 27th German Conference on Artificial Intelligence[C]. Ulm, Germany, 2004.

[37] Yuanchao liu, xiaolong wang, bingquan liu. A Feature Selection Algorithm For Document Clustering Based On Word Co-occurrence Frequency[A]. In: proceedings of the Third International Conference on Machine Learning and Cybernetics[C]. Shanghai, 26 - 29 August 2004.

[38] Z. Y. Niu, D. H. Ji and C. L. Tan. Document clustering based on cluster validation [A]. 13th Conference on Information and Knowledge Management[C]. CIKM 2004, 8 - 13 Nov 2004, Washington DC, USA.

[39] Stanislaw Osiński. Dimensionality Reduction Techniques for Search Results Clustering[D]. MSc. thesis, University of Sheffield, UK, 2004.

[40] Zhao, Y., Karypis, G. Criterion Functions for Document Clustering: Experiments and Analysis [A]. Technical Report #01—40, Department of Computer Science, University of Minnesota, Minneapolis, MN, 2001.

[41] Michael Steinbach, George Karypis, Vipin Kumar. A Comparison of Document Clustering Techniques [A]. Department of Computer Science and Engineering, University of Minnesota. Technical Report #00 - 034, 2000.