ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, Supplement, December 2009, pp.330−335

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

不完备文本信息系统的集对分析聚类算法*

林国平[1+], 李绍滋[2,3]

[1](漳州师范学院 数学系,福建 漳州 363000)

[2](厦门大学 智能科学与技术系,福建 厦门 361005)

[3](福建省仿脑智能系统重点实验室(厦门大学),福建 厦门 361005)

Clustering Method for Incomplete Text System Based on Set Pair Analysis

LIN Guo-Ping[1+], LI Shao-Zi[2,3]

[1](Department of Mathematics and Information Science, Zhangzhou Normal University, Zhangzhou 363000, China)

[2](Department of Cognitive Science, Xiamen University, Xiamen 361005, China)

[3](Fujian Key Laboratory of Brain-like Intelligent System, Xiamen University, Xiamen 361005, China)

+ Corresponding author: guoplin@163.com

Lin GP, Li SZ. Clustering method for incomplete text system based on set pair analysis. *Journal of Software*, 2009,20(Suppl.):330−335. http://www.jos.org.cn/1000-9825/09038.htm

Abstract: This paper presents a novel approach for incomplete text system. Which is based on hypergraph model clustering by using the set-pair analysis, and in which the similar, different and anti-contact connectivity of Set-pair and the similarity value of set-pair are used. After hypergraph model set up, a hypergraph partitioning algorithm is used to find clusters. This new method can eliminate disadvantageous factors and decrease the number of dimensions of the incomplete text data and enhance the speed largely and precision of text clustering. The experimental results show that the algorithm is feasibile and efficient.

Key words: incomplete system; set-pair analysis; high-dimensional clustering; hypergragh model; text clustering

摘 要: 考虑到实验数据的大规模性及不完备性等特点,根据集对分析理论,提出一种新超图模型不完备文本系统的聚类算法,即在超图边的权重中引入了集对的同异反联系度和集对的相似联系度并建立了超图模型,最后应用超图分隔法进行聚类.该算法克服了传统聚类算法的缺陷,更有效地降低了文本空间的维数,提高了不完备文本信息系统聚类的精度和速度.最后的实例说明了该算法的可行性和有效性.

关键词: 不完备信息系统;集对分析方法;高维聚类;超图模型;文本聚类

随着信息网络在全球范围内的兴起,信息处理的自动化是现今信息技术的研究热点.文本挖掘技术能降低网络查询时间,提高网络搜索质量,方便网络用户,能快速有效地获取文本信息.聚类分析作为一种数据挖掘的重要手段,在文本挖掘中也扮演着非常重要的角色.文本挖掘中聚类[1]是一个将文本集分组的全自动处理过程,是一种典型的无教师的机器学习问题.类是通过相关数据发现的一些组,类内的文本和其他组相比更为相近.因此,文本聚类的目标是找到这样一些类的集合,类之间的相似度最小,而类内部的相似性最大.即文本挖掘中的聚类分析是通过划分数据集发现潜在模式的知识挖掘过程,要求聚类结果簇间相似度最低而簇内相似度最高.

* Supported by the National Natural Science Foundation of China under Grant Nos.60873179, 10971186 (国家自然科学基金); the Foundation of Fujian Province Educational Department of China under Grant No.JB08187 (福建省教育厅B类基金项目)

Received 2009-05-03; Accepted 2009-09-30

而文本具有高维、稀疏的特征,所以此类数据聚类的难点在于[2−5]:(1) 距离函数难定义.聚类操作的基础是数据对象之间相似性的度量,相似度高的对象归为一类.低维空间中经常使用欧氏距离等距离函数来度量相似性,但在高维情况下由于相似性没有传递性,距离函数失效.必须重新考虑新的度量数据对象相似性的标准或准则.(2) 基于距离的聚类方法,经常需要计算簇的均值或近邻,但在高维情况下,按距离计算的簇的均值会很接近,聚类操作由于无法明确区分簇的中心而无法进行;(3) 由于维数很高,传统聚类算法的计算复杂度很高,其应用受到限制.

对于高维数据的处理,一种有效的方法是在保持数据关系上进行维规约,从而利用传统的聚类算法在低维的数据空间中完成聚类操作,如主成分分析(PCA)[6]、自适应*K*-means聚类(AKKC)[7]、自组织映射网络(SOM)[8]、基于概率模型的聚类分析技术[9]等,都是普遍应用的降维方法.类似PCA的潜在语义分析(LSI)[10]也是经常使用的降维技术.由于降维后,噪音数据与正常数据之间的差别缩小,由此得到的聚类结果质量较差.另外,降维技术的使用虽然缩小了数据维度空间,但其可解释性、可理解性较差,可能会丢失重要的聚类信息,其结果的表达和理解也存在着一定的难度.

而在数学界,图论已被证明是解决运筹学和优化领域中重要问题非常有用的工具.为了解决更多的组合问题,把图的概念进行推广是非常自然的事情.图概念是Berge于 1970 年提出的.由于超图理论比较抽象,研究者很不容易入门,超图理论的发展一直比较缓慢.近年来,超图理论及其应用的研究越来越为人们所重视.基于超图模型的分割可能适用于高维数据的聚类,高维空间的关系转化成超图,用超边的权重来描述空间点间的关系.对超图的分割实际上就是聚类的过程,将权重大的超边中包含的数据点尽量放在一个类中,同时使被切割的超边权重之和最小[11].

本文根据集对分析理论,提出一种新超图模型聚类算法,即在超图边的权重中引入了集对的同异反联系度和集对的相似联系度并建立了超图模型,最后应用超图分隔法进行聚类.该算法克服了传统聚类算法的缺陷,即该类算法只能有效地应用于空间维数较小的情况,对于高维空间则不能产生有意义的聚类结果.而本文提到的算法更有效地降低了文本空间的维数,提高了文本聚类的精度.

1 集对分析法的有关概念

集对分析(set-pair analysis,简称SPA)[12]方法是由赵克勤教授近年来提出的用于研究集合之间相互关系的一种新理论,其核心思想是把被研究的客观事物之确定性联系和不确定性联系作为一个系统来分析处理,所谓集对就是具有一定联系的两个集合所组成的基本单元,集对加以刻画,进而得出这两个集合在所论问题背景下的关心程度的表达式.通过这一思想,我们就可以进行不完备信息系统的相似度的比较.现在已经得到广泛的应用.下面介绍相关的几个定义.

定义 1[12]**(同异反联系度)**. 设集合 M_1,M_2 组成集对 $H(M_1,M_2)$,根据问题背景对集对H的属性展开分析.假设有n个属性,其中有s个为集对H中的两个集合所共有的属性,p个为集对H中两个集合所相互对立的属性,剩下的 $f=n-s-p$ 个属性既不对立,又不同一,即其性质的关系不能确定,则称 s/n 为这两个集合的同一度, f/n 为两个集合的差异度, p/n 为两个集合的对立度,并用公式:

$$\mu(M_1,M_2)=\frac{s}{n}+\frac{f}{n}i+\frac{p}{n}j=a+bi+cj \tag{1}$$

表示集对的联系度,我们称μ为集对的同异反联系度表达式.式(1)中 j 为对立度系数,规定 $j=-1$,表示 c 是与 a 方向相反的量.不计 j 的值,仅作为反向量的标记使用.i 为差异度系数,在[−1,1]区间视不同情况取值,以说明 b 处在 a 与 c 的某个中间位置上.不计 i 的值时,仅作为异分量的标记使用.

显然,在定义 1 中,a,b,c 这 3 个数满足归一化条件 $a+b+c=1$ 且 $0\le a,b,c\le 1$.其中,

① $a>>c$ 表示集对的同一度越大;反之,相反程度越大;

② $a>>b$ 表示集对同一可能性越大;反之,表示不确定性程度越高,同一度越小;

③ $c>>b$ 表示集对对立可能性越大;反之,集对不确定性程度越高,对立度越小.

定义 1 认为集对中的各个属性量纲是相等的,而实际情况分析出的集对属性应该是不均匀的,所以我们对分析出的属性赋予权值 w_i,且使 $\sum_{i=1}^{n} w_i = 1$,通过权值来表示各属性在集对相似性计算中所占的比重.

定义 2(集对属性度). 设(U,A)是一个文本信息系统,其中 U 是对象集,A 是属性集. $\forall x,y \in U$, $a(x) \in V$,V 是属性值域,定义 x 与 y 的集对 $H(x,y)$ 属性集如下:

(1) $M(x,y) = \{a \in A / a(x) \neq * \wedge a(y) \neq * \wedge a(x) \neq a(y)\}$ (取值都明确但相反的属性个数);

(2) $N(x,y) = \{a \in A / a(x) \neq * \wedge a(y) \neq * \wedge a(x) = a(y)\}$ (取值都明确且相同的属性个数);

(3) $Q(x,y) = \{a \in A / a(x) = * \vee a(y) = * \vee a(x) = a(y)\}$ (取值都明确但既不相同又不相反的属性个数),

其中,"*"是表示未知属性值.

定义 3[13](集对的同异反联系度). 设集对$H(x,y)$中的对象x,y各有n个属性,各属性权重分别表示为 $w_x(a_i), w_y(b_i), i = 1,2,...,n$ 且 $\sum_{i=1}^{n} w_x(a_i) = \sum_{i=1}^{n} w_y(b_i) = 1$,其中有$m_1$对属性是相同的,权重表示为 $w(a_i)$;m_2对属性是相反的,权重分别表示为 $w(c_i)$ 和 $w(C_i)$; 所以得到 $n - m_1 - m_2$ 对属性是既不相同也不相反的,权重分别表示为 $w(b_i)$ 和 $w(B_i)$,在这种情况下,定义 x,y 同异反联系度表达式:

$$\mu = a + bi + cj \tag{2}$$

其中, $a = \frac{1}{l}\sum_{i=1}^{m_1} w^2(a_i)$, $c = \frac{1}{l}\sum_{i=1}^{m_2} w(c_i).w(C_i)$, $b = 1 - a - c$, $l = \sqrt{\sum_{i=1}^{n}(w_{x_i})^2}.\sqrt{\sum_{i=1}^{n}(w_{y_i})^2}$ 且 $m_1 = |M(x,y)|, m_2 = |N(x,y)|$, $n - m_1 - m_2 = |Q(x,y)|$.

定义 4(集对联系度). 设(U,A)是一个文本信息系统,其中 U 是文本对象集,A 是属性集. $\forall x,y \in U$ 定义 x 与 y 的集对联系度 $U(x,y)$ 为一个三元组; $U(x,y) = (s_0, s_1, s_2) = s_0 + s_1 i + s_2 j$; 这里 $s_0 = |M(x,y)| / n$ 称为 x,y 同一度, $s_1 = |N(x,y)| / N$ 称为对立度, $s_2 = |Q(x,y)| / N$ 称为差异度;其中 $N = |A|$. 显然, $s_0 + s_1 + s_2 = 1$.

2 基于集对分析的超图模型文本聚类算法

2.1 文本的表示

如何使文本易于被计算机处理,是文本挖掘所面临的前期工作,近年来这方面的研究工作已经取得了一定的进展.目前应用较多且效果较好的是向量空间模型(vector space model,简称 VSM)法.

在向量空间模型中,文本空间被看作是由一组正交词条向量所组成的向量空间,每个文本表示为其中一个范化特征向量 $V(d) = (t_1, w_1(d); t_2, w_2(d); ...; t_n, w_N(d))$,其中$t_i$为词条项, $w_i(d)$ 为t_i在文本d中的权重. $w_i(d)$ 一般被定义为t_i在文本d中出现频率 $tf_i(d)$ 的函数,即 $w_i(d) = \phi(tf_i(d))$ 在VSM中,TF-IDF(term frequency inverse document frequency)是一种常用的词条权重确定方法.TF-IDF的计算公式:

$$W_i(d) = tf_i(d) \cdot \log(N / n_i) \tag{3}$$

其中,N为所以文本数目,n_i为含有词条t_i的文本数目.由于t_i在文本中既可以重复出现又应该有先后次序关系,分析起来有一定难度,为了简化分析,可以暂不考虑t_i在文本中的先后次序并要求t_i互异(即没有重复).这时可以把 $t_1, t_2, ..., t_n$ 看成是一个n维的坐标系,而 $w_1, w_2, ..., w_n$ 为相应的坐标值,因此一个文本就表示为n维空间的一个向量,我们称 $V(d) = (w_1, w_2, ..., w_N)$ 为文本d的向量表示.

我们可以这样理解向量空间中的每一个分量;每个分量刻画了项t_i区分文本内容属性的能力,一个项在文本集中出现范围越广,说明它区分文本内容属性的能力越低;在一个特定文本中出现频度越高,说明它在区分该文本内容属性方面的能力越强.

事实是,由网页特征向量组成的集对也可以看作一组不完备信息.特征向量中任意两个属性之间可能存在语义上的相同、相反或不确定的关系.分析由网页特征向量组成的集对时,我们可以通过语义确定两个向量中

关键字是否属于同一类型属性,在该项属性中两关键字是同一关系、相反关系还是相异关系.比如假设两个文本特征向量分别为 $V(d_1)$ =(电脑,0.5;红色,0.3;桌子,0.6;广告,0.2); $V(d_2)$ =(pc,0.2;黄色,0.5;三角架,0.3).在我们分析这两篇文档的相似度时,可以得到电脑和pc是同一类属性,语义上是同一关系;红色和黄色是一类属性,语义上是相反关系,桌子和三角架是一类属性,语义上是相异关系;d_1中广告属性在d_2中找不到对应项,也作为不确定属性.这样就形成了一个带有不完备信息的集对.

2.2 文本相似度的计算

假设有 N 个文本对象为 $\{d_1,d_2,...,d_N\}$,描述第 k 个对象的 n 个词条的权重分别为 $w_{ki},i\in\{1,2,...,N\}$,且已知集对 $H(d_i,d_j)$ 的同异反联系度 $\mu(d_i,d_j)=a+bi+cj$ 和集对联系度 $U(d_i,y_j)=s_0+s_1i+s_2j$,则文本 d_i,d_j 相似度(属性分布相似度)定义为

$$Sim(d_i,d_j)=p\times q\times l \tag{4}$$

其中 $p=c-a,q=b-a,l=c-b$ 规定 $(Sim)_1\prec(Sim)_2$ 等价于

$$(Sim)_1\prec(Sim)_2=\{(p_i\prec p_j)\,||\,(p_i=p_j\,\&\&\,y_i\prec y_j)\,||\,(p_i=p_j\,\&\&\,y_i=y_j\,\&\&\,l_i\prec l_j)\}.$$

Sim 越大,表示数据对象越相似.越可能被聚类到同一簇中.

2.3 算法描述

步骤 1.建立超图模型.

所谓数据集超图模型指的是将 n 个数据对象映射为高维空间中的 n 个点(在不引起混淆的情况下,该文不区分“数据对象”和“点”),并对其建立超图模型 $H=(V,E)$,其中,$V=\{v_1,v_2,...,v_N\}$ 是超图的顶点集合,$E=\{e_1,e_2,...,e_t\}$ 超图模型的边集合(设共有 t 边).超图是图的扩展,其中每条边都可连接两个以上的结点.本文所建的超图模型是:假定文本信息系统 (U,A) 共有 N 条记录,每条记录对应一篇文档的特征向量,每个特征向量含有 n 个词项分量(视为属性),每个属性权重由公式(3)计算得到,则 N 个特征向量对应超图中 N 个顶点,第 i 个顶点的坐标为 n 个属性值,建立超图的过程,实质是寻找二元组超边的过程 $e_i=\{V,w_i\}$,其中 $V=\{v_1,v_2,...,v_p\}$ 是顶点集合(此处每条超边连接的结点数 p 恒为 2),w_i 表示相应边的权重.根据预先给定的某一阈值,在相似性大于该值的两点间建立超边,得超边集合 $E=\{e_1,e_2,...,e_m\}$,超边的权重等于两点间属性分布相似度.超边对应的是点与点之间的相似性,相似性越高,边的权重就越大.

步骤 2.聚类.

基于超图模型的聚类算法可以有多种,如模拟退火优化算法[14]、超图分割算HMETIS[15]等.应用图的分割算法对图进行切割,基本思想是:在图中寻找权重最小的边并将它截断,不断反复直到得到k个分散的超图子集.这k个超图的分支,其内部结点之间的相似性最强,即图内的连通度强,形成数据簇.超图分割算法HMETIS[10],即是依据此思想,在超图模型基础上每次将超图分成两部分,并保证被截断的超边的权重最小.超边的权重越小,说明超边表示的关系越不重要,反复使用分割算法,直到每个分割内部都紧密联系为止,得到的分割就是簇.

假定,依据超图聚类算法已得到k个分隔 $G=\{G_1,G_2,...,G_k\}$,需要对聚类结果进行评价,以确定所得聚类过程是否继续.对聚类结果进行评价,一般通过两个指标进行,一是簇之间相似性,一是簇内相似性.在此,这里主要考察簇(即超图子集)内部数据结点的集对联系度.由集对联系度 $U(x,y)$中的同一度,对立度和差异度的大小可判断超图子集是否继续分割.

3 算法分析

该文提出的算法只需遍历数据集 1 次,对数据的输入顺序不敏感.此外,与其他基于超图模型的传统聚类算法相比,还有以下优点:

(1) 模型建立过程简单:该文通过集对的同异反联系度和集对的相似联系度来定义文本间的相似度,并用之来直接计算两点之间属性分布相似度.

(2) 应用形式多样:针对二元数据,该文算法可直接应用进行聚类分析;对于连续值或字符值数据,也可以作

为数据预处理算法,首先对文本的属性分布相似度进行度量,并在得到较小数据集上应用传统算法进行低维聚类.

(3) 从计算复杂度的角度分析,如果恰当的选择相似度阈值,对于一个拥有上百万个文本的数据库,使用该算法几分钟内就能完成聚类.显然,阈值如果设置太低,将使计算复杂度成倍上升,但是如果设置过高,将出现太多的孤立点,这样可能漏掉一些很重要的信息,导致聚类质量的降低.

4 实验结果

为了评价基于集对分析的超图模型的文本聚类算法和传统聚类算法的相对性能,我们从人民日报上共抽取 260 篇文章,其中经济类文章 90 篇,政治类文章 85 篇,教育类文章 85 篇,这些文章应用 TF-IDF 方法经过分词处理和特征选取以后,得到文本向量的维数 2 437 维,则构造 2437×260 维的向量空间,采用查全率和查准率来评价这些聚类结果.分别按照传统聚类算法如文献[9]中的算法和基于集对分析的超图模型的文本聚类算法对测试文档集进行独立的测试,记录其聚类结果,并比较两种算法的性能.实验结果表明了基于集对分析的超图模型的文本聚类算法在文本聚类处理中的可行性.

实验中,相关参数分别为:相似度阈值设为 0.05,超图划分的分支数 k=64.即构造超图模型时,相似度小于 0.05 的两个点没有边,超图模型共包括 512 个顶点和 52 962 条超边.由于一些文本间的相似度过小,从而作为孤立点处理,因此超图的顶点数小于文本总数.可见,最小阈值提供了一个简单而有效的从超图模式中去除不必要信息的方法.

Table 1 Clustering result of HMETIS algorithm

表 1 利用文献[9]算法得到的聚类结果

结果对比	文本聚类结果							
	经济类	政治类	教育类	同时属于经济和政治	同时属于经济和教育	同时属于政治和教育	同时属于经济和政治和教育	正确率(%)
经济类(90)	71	14	5	0	0	0	0	78.9
政治类(85)	9	70	6	0	0	0	0	82.4
教育类(85)	13	3	69	0	0	0	0	81.2
平均正确率(%)	81.6							

Table 2 Clustering results of the algorithm of this paper

表 2 利用本文算法的聚类结果

结果对比	文本聚类结果							
	经济类	政治类	教育类	同时属于经济和政治	同时属于经济和教育	同时属于政治和教育	同时属于经济和政治和教育	正确率(%)
经济类 90	41	1	3	31	12	2	1	93.3
政治类 85	4	37	0	31	2	10	3	92.9
教育类 85	3	1	31	3	11	28	1	91.8
平均正确率(%)	92.7							

实验结果表明,利用文献[9]中算法聚类平均正确率为 81.6%,基本上达到聚类的效果,而利用本文的聚类算法聚类的平均正确率为 92.7%.并且有些文章同时属于多个类,经过阅读这些文本,发现聚类结果与实际情况相吻合.由此可以看出,基于集对分析的超图模型的文本聚类算法在文本自动聚类处理是可行的;且聚类速度和效率均明显优于文献[9]中算法的速度和效率.由实验可知,当给定合适的阈值后,数据集中绝大多数数据对象(即超图结点)的关系都能在超图模型中得到反映,少数噪声数据或奇异点或者在建模时未被纳入图中结点,或者在超图分割过程中成为孤立点,灵活性高,数据分析全面;而采用其他超图方法往往容易丢失数据,而丢失的数据也许包含重要的信息.

5 结 论

高维文本聚类是数据挖掘技术中的一个重要课题.本文提出了基于集对分析的不完备文本信息系统的聚

类算法,避免了有些传统聚类的较大时间复杂度,且使它的适用性更广,更能体现一个聚簇的规律,保证了文本聚类质量.通过分析可以看到该算法在文本聚类处理中的可行性.

正像其他聚类算法一样,该文提出的聚类算法也存在如何选择合适的聚类参数的问题,有两个关键的参数需要确定,一是集对相似度阈值,二是超图划分的分支数.因此如何确定合适的参数是我们未来工作内容之一.

致谢 在此,作者向对本文的工作给予支持和建议的同行,尤其对论文评审专家表示衷心感谢.

References:

[1] Li XG, Yu G, Wang DL, Bao YB. Latent concept extraction and text clustering based on information theory. Journal of Software,2008,19(9):2276−2284 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2276.htm

[2] Bennett K, Ferris M, Ioannidis YE. A genetic algorithm for database query optimization. In: Proc. of the 4th Int'l Conf. on Genetic Algorithm. Morgan Koffmann Publishers, 1991. 400–407.

[3] Mörchen F, Brinker K, Neubauer C. Any-Time clustering of high frequency news streams. In: Data Mining Case Studies Workshop, the 13th ACM SIGKDD, Int'l Conf. on Knowledge Discovery and Data Mining. 2007.

[4] Geng L, Hamilton HJ. Interestingness measures for data mining: A survey. ACM Computing Surveys, 2006,38(3).

[5] Malik HH, Kender JR. High quality, efficient hierarchical document clustering using closed interesting itemsets. In: Proc. of the 6th IEEE Int'l Conf. on Data Mining. 2006. 991–996.

[6] Jackson JE. A User's Guide to Principal Components. John Wiley & Sons, 1991.

[7] Liu XH, Yu S, Moreau Y, De Moor B. Hybrid clustering of text mining and bibliometrics applied to journal sets. In: Proc. of the SIAM Data Mining Conf. 2009.

[8] Kohonen T. Self-Organization and Associated Memory. Springer-Verlag, 1998.

[9] Jiang N, Shi ZZ. Bayesian posteriorimodel selection for text clustering. Journal of Computer Research and Development, 2002, 39(5):341−346 (in Chinese with English abstract).

[10] Karypis G, Kumar V. A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM Journal on Scientific Computing,1998,20(1):359–392.

[11] Chen YK, Lu ZD, Guo J. A high dimension clustering algorithm and its application in detecting money laundering. Computer Science, 2007,34(6):191−213 (in Chinese with English abstract).

[12] Zhao KQ. Set Pair Analysis and its Preliminary Application. Hangzhou: Zhejiang Science Press, 2000. 50−55 (in Chinese).

[13] Ma SM, Wang RC, Ye N. An uncertainty reasoning approach based on set pair analysis for context awareness. Journal of Nanjing University of Posts and Telecommunication, 2009,29(1):64−67 (in Chinese with English abstract).

[14] Kirkpatrick S, Gelatt CD, Vecchi HMP. Optimization by simulated annealing. Science, 1983,220(4598):671–680.

[15] Karypis G, Aggarwal R, Kumar V, *et al*. Multilevel hypergragh partitioning application in VLSI design. In: Proc. of the ACM/IEEE Design Automation Conf. Anaheim: ACM Press, 1997. 526–529.

附中文参考文献:

[1] 李晓光,于戈,王大玲,鲍玉斌.基于信息论的潜在概念获取与文本聚类.软件学报,2008,19(9):2276−2284. http://www.jos.org.cn/1000-9825/19/2276.htm

[9] 姜宁,史忠值.文本聚类中的贝叶斯后验模型选择方法.计算机研究与发展,2002,39(5):341−346.

[11] 陈云开,卢正鼎,刘芳,郭洁.一种高维聚类算法及在洗钱侦测中的应用.计算机科学,2007,34(6):191−213.

[12] 赵克勤.集对分析及其初步应用.杭州:浙江科学出版社,2000.50−55.

[13] 马守明,王汝传,叶宁.基于集对分析的上下文感知不确定性的推理方法.南京邮电大学学报(自然科学版),2009,29(1):64−67.

林国平(1978－),女,福建福安人,讲师,主要研究领域为数据挖掘,人工智能.

李绍滋(1963－),男,教授,博士生导师,主要研究领域为人工智能与多媒体信息检索,计算机视觉与机器学习,网络多媒体及CSCW 技术.