粗糙集理论中新的针对不完备信息系统的处理方法研究*)

潘 巍[1] 王阳生[2] 杨宏戟[3]
(首都师范大学信息工程学院 北京 100037)[1]
(中国科学院自动化研究所模式识别国家重点实验室 北京 100080)[2]
(Software Technology Research Laboratory, De Montfort University, Lercester, LEI19BH, England)[3]

摘 要 本文针对不完备信息系统,在分析了现有的数据补齐法和扩充法的优劣后,提出基于差异关系和数据部分补齐的处理方法。本文定义了差异关系,对差异矩阵进行扩充使其能适用于不完备信息系统,证明了用差异关系进行属性约简和求核的可行性,并给出了相应的算法。数据部分补齐的目的是通过分析对象之间的差异性更进一步地挖掘信息之间的潜在联系,为后续的决策规则的求取提供更丰富更准确的信息。实验证明,在处理不完备信息系统时,基于差异关系和数据部分补齐的方法能获得更好的分类性能。

关键词 粗糙集,不完备信息,相似关系,数据补齐

Research on Rough Sets under Incomplete Information System

PAN Wei[1] WANG Yang-Sheng[2] YANG Hong-Ji[3]
(Institute of Information Engineering, Capital Normal University, Beijing 100037)[1]
(Institute of Automation, Chinese Science Academies, Beijing 100080)[2]
(Software Technology Research Laboratory, De Montfort University, Leicester, LE1 9BH, England)[3]

Abstract In order to process incomplete information systems in Rough sets, a novel method based on discrepancy relation and data partial supplement is introduced. This paper analyzes current typical methods, defines the discrepancy relation, extends the discrepancy matrix, proves the feasibility of using discrepancy relation to reduce redundant attributes and provides corresponding algorithms. Besides, uses a new method of data partial supplement to obtain latent data to provide more accurate information for next step of extracting decision rules. The experiment results show that the new method can get best performance evaluation.

Keywords Rrough set, Incomplete information, Similarity relation, Data recruitment

1 引言

粗糙集理论是波兰数学家 Z. Pawlak 在 1982 年提出的一种数据分析理论,由于对不确定性的描述相对客观,目前,该理论已被成功地应用于机器学习、决策分析、模式识别、数据挖掘等领域,并取得了成功。

遗憾的是粗糙集理论是基于完备系统这样一个假设,即信息表中所有的属性值都是已知的。而在实际应用中,由于数据的测量误差、对知识获取的限制等原因,人们往往面临的是不完备信息系统,即可能存在部分属性值未知的情况,此时,直接应用原有的粗糙集理论往往不能得到正确的结论。因此,如何有效地将粗糙集理论应用于不完备信息系统,一直是学者研究的主要方向之一。

2 粗糙集理论中新的针对不完备信息系统的处理方法

为了使粗糙集理论能够适应不完备信息系统的处理,目前主要有两大类方法:一是数据补齐法,一是扩充法。

2.1 现有数据补齐法简介

数据补齐法[1~4],即通过某种方法(例如概率统计、可能值穷举等)填补所有未知的属性值,从而把不完备信息系统转化为完备信息系统,然后再进行知识约简。数据补齐法可分为几个大类:概率法、基于熵法、穷举法、忽略法和特定值法。

数据补齐可以说是一种间接处理方法,补齐后的信息表可能与实际情况有出入,为保证知识约简后规则的准确性和可靠性,数据补齐原则的选取十分重要。Grzymala-Busse 等[1]使用 10 个不同的数据库比较了 9 种数据补齐法,发现在不同的评价标准下,各种数据补齐法的性能虽有所不同,但同时考虑条件属性与决策属性的方法的性能要优于只考虑条件属性的方法。

2.2 现有扩充法简介

扩充法,即针对粗糙集理论中的相关概念在不完备信息系统中进行适当扩展,使系统能够直接从不完备信息表中进行知识约简和求核,得到所需的融合算法。相对于数据补齐法而言,扩充法是一种直接处理方法。典型的扩充法有:基于容差关系的扩充模型[5]、基于相似关系的扩充模型[6]、基于量化容差关系的扩充模型[6,7]和基于限制容差关系的扩充模型[8]等。

现有的几种扩充模型中,容差关系的条件最宽松,只要两个个体有相似的可能,就会被分到同一个容差类,例如:{ *,

*)基金项目:国家 863 高技术研究发展计划项目(编号:2003AA114020);潘 巍 博士,讲师,研究方向为模式识别,信息融合。王阳生 研究员,博士生导师,研究方向为模式识别。

1,*,*}和{*,*,2,*}。相似关系过于苛刻,可能将具有很多相同已知属性信息的个体分到不同的相似类,例如:{1,*,3,4,5}和{*,2,3,4,5}。量化容差关系模型虽然可以刻画两个对象之间的相似程度,但需要预先知道信息系统中属性值的概率分布情况,而这对于一个新的不完备信息系统来说是很困难的。限制容差关系模型则对容差类的相似程度加以约束,要求两个对象必须拥有至少一组相同的已知属性信息,因此分类效果介于容差关系和相似关系这两个极端情况之间。

2.3 基于差异关系的扩充法

1. 差异关系及相关概念的定义

本文认为,不论是容差关系、相似关系,亦或是它们的变体,其最终目的都是为区分两个对象提供依据。事实上,相似关系、量化容差关系及限制容差关系都可以看作是在容差关系的基础上对相似性进一步细分,只不过用于判断是否相似或相似程度的标准有所差异。换个角度思考,如果用相似性进行区分会产生分歧,那为什么不改用差异性呢?只要两个对象具有不同的属性值,我们就可以很明确地区分它们。所以,在不完备信息系统中,与其求取对象之间的相似性,不如直接考虑对象之间的差异性,二者殊途同归,都是对各个对象进行划分,但基于差异的划分是明显的,而基于相似的划分却有多种形式。据此,本文提出基于差异关系的扩充模型。

定义 1 差异关系 DR 的定义为:

$$\forall x,y\in U,\forall B\subseteq A,\text{有 } DR_B(x,y)\Leftrightarrow c\in B,(c(x)\neq c(y))\wedge c(x)\neq *\wedge c(y)\neq * \tag{1}$$

令 $I_B^{DR}(x)$表示在属性集合 B 上满足关系 $DR_B(x,y)$的对象 y 的集合,即 x 的差异类,令 $\overline{I}_B^{DR}(x)=U-I_B^{DR}(x)$表示对象 x 的非差异类。

采用差异性判断要比相似性判断容易得多,对于对象 x 和 y,在进行比较时,如果发现二者有一组属性值不相同,即可立刻判断 x 和 y 为差异类,而用相似性进行判断时,可能还需要考虑更多的因素。此外,差异类也没有计算上、下近似集的必要,大大减少了计算量。

定义 2 异决策差异关系 DDR 的定义为:

$$\forall x\in U,\forall B\subseteq A,\text{有 } DDR_B(x,y)\Leftrightarrow y\in I_B^{DR}(x)\wedge d\in D,d(x)\neq d(y) \tag{2}$$

令 $I_B^{DDR}(x)$表示在属性集合 B 上满足关系 $DDR_B(x,y)$的对象 y 的集合,即 x 的异决策差异类。显然,$I_B^{DR}(x)-I_B^{DDR}(x)$就是同决策差异类。

定义 3 对象 x 的异决策非差异关系 DSR 的定义为:

$$\forall x\in U,\forall B\subseteq A,\text{有 } DSR_B(x,y)\Leftrightarrow y\in\overline{I}_B^{DR}(x)\wedge d\in D,d(x)\neq d(y) \tag{3}$$

令 $I_B^{DSR}(x)$表示在属性集合 B 上满足关系 $DSR_B(x,y)$的对象 y 的集合,即 x 的异决策非差异类。显然,$I_B^{DR}(x)-I_B^{DSR}(x)$就是同决策非差异类。

获得对象 x 的差异类、异决策差异类、异决策非差异类后,就可以根据需要将其直接用于下一步的计算。

2. 扩充差异矩阵的定义

在经典粗糙集理论中,波兰数学家 Skowron 提出用 Skowron 矩阵(也称差异矩阵)来描述一个信息系统中各对象之间的属性差异,它是进行知识约简的有利工具。本文在差异关系的基础上对其在不完备信息系统下进行扩充,并称其为扩充差异矩阵。

定义 4 给定不完备信息系统 $S=\langle U,A,V,f\rangle$,$|U|=n$,基于差异关系的扩充差异矩阵 M_{xy}是一个 $n\times n$ 的矩阵,其任一元素 $m(x,y)$是满足 $y\in I_A^{DR}(x)$的属性 $a\in A$ 的集合。如果 $y\notin I_A^{DR}(x)$,则 $m(x,y)=\Phi$。相应地,给定不完备信息系统 $S=\langle U,C\cup D,V,f\rangle$,$|U|=n$,则基于差异关系的决策扩充差异矩阵 DM_{xy}是一个 $n\times n$ 的矩阵,其任一元素 $dm(x,y)$是满足 $y\in I_C^{DDR}(x)$的属性 $c\in C$ 的集合。如果 $y\notin I_C^{DDR}(x)$,则 $dm(x,y)=\Phi$。

有的文献倾向于使用二值化的差异矩阵,通过差异关系同样能很方便地获得。使用二值化的差异矩阵前,首先需要生成差异比较表。差异比较表通常是一个 $n\times(n-1)/2$ 行 n 列的矩阵($n=|U|$),矩阵的每一列对应一个属性,每一行对应一对不同的对象。

定义 5 给定不完备信息系统 $S=\langle U,A,V,f\rangle$,$|U|=n$,基于差异关系的二值化扩充差异矩阵 BM_{xy}是一个 $n\times(n-1)/2$ 行 n 列的矩阵,对任一元素 $bm((x,y),i)$,

$$bm((x,y),i)=\begin{cases}1, & y\in I_A^{DR}(x)\wedge A_i(x)\neq A_i(y)\neq *\\ 0, & \text{else}\end{cases} \tag{4}$$

相应地,给定不完备信息系统 $S=\langle U,C\cup D,V,f\rangle$,$|U|=n$,则基于差异关系的二值化扩充决策差异矩阵 BDM_{xy}是一个 $n\times(n-1)/2$ 行 n 列的矩阵,对任一元素 $bdm(x,y)$,

$$bdm((x,y),i)=\begin{cases}1, & y\in I_A^{DDR}(x)\wedge A_i(x)\neq A_i(y)\neq *\\ 0, & \text{else}\end{cases} \tag{5}$$

扩充差异矩阵可用于设计不完备信息系统的知识约简算法,实现对不完备信息系统的直接处理。

2.4 基于差异关系和数据部分补齐的不完备信息系统处理方法

余英泽等[8]在 Hayes 数据集上使用扩充法与数据补齐法分别进行分类识别的比较结果显示,扩充法和数据补齐法的正确率互有高低,没有绝对的强弱。本文认为,数据补齐法与扩充法各有优劣,扩充法善于寻找知识之间的内在联系,数据补齐原则可以提供知识属性的表层分析,二者的结合能更好地挖掘知识的潜在规律,提高决策规则的正确性和可靠性。据此,本文提出了基于差异关系和数据部分补齐的不完备信息系统处理方法。

2.4.1 基于差异关系的属性约简和求核

目前国内外关于属性约简的算法多是基于完备信息系统的,研究的重点是探讨更快速更高效的属性约简算法及全面的或最优的属性约简表达式,如果将其直接应用于不完备信息系统,有时并不能得到令人满意的结果。对此,最简单的思路就是使用扩充模型使属性约简算法能够适用于不完备信息系统。由于本文提出的基于差异关系的扩充差异矩阵具有生成方便、不需要计算上、下近似集等优点,因此可以直接与属性约简算法相结合进行计算。

这里,本文给出一个根据差异关系快速计算核属性的方法。

给定不完备信息表 $S=\langle U,C\cup D,V,f\rangle$,可以获得关于 S 的完备信息子表 S_0,S_0 中存放的是 S 中具有完备信息的所有对象。

定理 1 给定不完备信息表及其完备信息子表 $S=\langle U,C\cup D,V,f\rangle$及其完备信息子表 S_0,$c_i\in C$,如果 c_i 是 S_0 的核属性,则它一定也是 S 的核属性。

证明:如果 c_i 是 S_0 的核属性,则表明去除 c_i 后,$\exists x,y\in U_0,\forall b\in C,b(x)=b(y)\wedge d\in D,d(x)\neq d(y)$,$S_0$ 中会产生决

策矛盾，而$U_0 \subseteq U$，表明在S中，如果去除c_i，同样会产生决策矛盾，因此c_i也是S的核属性。证毕。

定理2 给定不完备信息表$S=\langle U, C \cup D, V, f\rangle$，$B \subseteq C$，$I_B^{DR}(x)$是对象$x$的差异类。对于$b \in B$，$b$是$B$中必要的属性，如果$\exists y \in U, y \notin I_{B-\{b\}}^{DR}(x) \wedge y \in I_B^{DR}(x)$。

证明：设$x, y \in U, x \neq y$，根据属性集B获得的$I_B^{DR}(x)$中包含y，而根据属性集$B-\{b\}$获得的$I_{B-\{b\}}^{DR}(x)$中不包含y。这表明，属性b可以分辨x和y，即有$b(x) \neq b(y) \wedge b(x) \neq * \wedge b(y) \neq *$，但$B$中除$b$外的其它属性均不足以分辨$x$和$y$，$x$和$y$中只有一对互不相同的已知属性值，它是区分$x$和$y$的唯一凭据，所以$b$是$B$中不可约简的属性。证毕。

例1 以Stefanowski在文[6]中提供的不完备信息表为例，用差异关系求取核属性。表1中，U是对象集合，c1～c4是4个条件属性，d是决策属性。此信息表是不协调的，因此在计算核属性前，要将其转为协调的信息表。在这里，我们直接删除a2和a3，经计算，(c1, c2, c4)是核属性。

利用差异关系获得不完备信息表的核属性并得到需要的属性约简后，就可以进入下一步的计算：基于差异关系的数据部分补齐。

表1 利用差异关系求取核值的不完备信息表

U	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}
c1	3	2	2	*	*	2	3	*	3	1	*	3
c2	2	3	3	2	2	3	*	0	2	*	2	2
c3	1	2	2	*	*	2	*	0	1	*	*	1
c4	0	0	0	1	1	1	3	*	3	*	*	*
D	Φ	Φ	Ψ	Φ	Ψ	Ψ	Φ	Ψ	Ψ	Φ	Ψ	Φ

2.4.2 基于差异关系的数据部分补齐

本文认为，如果能够从现有的无缺失值的样本对象中分析出某些确定的规则，那么，进行数据部分补齐会提高最终的识别效果，但这主要是针对具有少量未知属性的样本对象进行，补齐时还应考虑样本对象间的差异关系。

定理3 决策表$S=\langle U, C \cup D, V, f\rangle$，$B \subseteq C$，$x \in U$，$x$中有缺失值，$I_B^{DR}(x)$表示在属性集合$B$上$x$的差异类，$\overline{I}_B^{DSR}(x)$表示在属性集合$B$上$x$的异决策非差异类。则$x$进行数据补齐后，不会与$I_B^{DR}(x)$产生决策矛盾；而只可能会与$\overline{I}_B^{DSR}(x)$产生决策矛盾。

证明：1)根据差异类的定义，在对象x的差异类$I_B^{DR}(x)$中，$\forall y \in I_B^{DR}(x)$，必然存在至少一个属性$b \in B$，$b(x) \neq b(y) \wedge b(x) \neq * \wedge b(y) \neq *$，而决策矛盾是指决策表中由两个对象得到的决策规则前件相同，但后件不同。因此，x在补齐后不会与$I_B^{DR}(x)$产生决策矛盾。

2)针对非差异类，设$z \in \overline{I}_B^{DR}(x)$且$x$和$z$在数据补齐后形成决策矛盾，即$x$和$z$的前件相同，后件不同，其前件就是$\forall b \in B, b(x)=b(z)$，后件即是$\exists d \in D, d(x) \neq d(z)$，显然，唯一能同时满足前件和后件要求的就是具有异决策的非差异类$\overline{I}_B^{DSR}(x)$，即对象x在数据补齐后只可能与其异决策非差异类$\overline{I}_B^{DSR}(x)$产生决策矛盾。证毕。

根据定理3，对象x进行数据补齐后，只可能与其异决策非差异类$\overline{I}_B^{DSR}(x)$产生决策矛盾，而不会与非$\overline{I}_B^{DSR}(x)$类的对象产生决策矛盾。由此，本文提出基于差异关系的数据部分补齐，即排除会产生决策矛盾的属性值，减少缺失值的可能取值范围，如果可能取值数只有1个，那么就可以确定缺失值的属性值，并进行补齐。为确保数据的可靠性，本文仅对具有一个缺失值的对象进行数据补齐的运算。

例1中，a_{12}的$\overline{IC}_{new}^{DSR}(a_{12})=\{a_9\}$，为避免决策矛盾，可断定$a_{12}$的$c_4$属性值必不为3，在已知信息表$S$中，可选值1或0，将其记入$PossibleAttValueof a_{12}(c_4)$集合备用。此时，如果$a_1$的决策属性是Ψ，那么可以确定$a_{12}$中$C_4$应取值为1，$a_{12}$变成完备信息，并记入$S_{new0}$。

基于差异关系的数据部分补齐的目的不仅能避免决策矛盾的产生，更可为下一步的计算提供更准确更完备的信息，因此，在属性值域较小，不完备信息表中具有完备信息的对象又相对多时，在提取决策规则前进行数据部分补齐是有益的。但当属性值域较大，具有完备信息的对象又较少时，有时并不能得到预期的效果，通常，不完备信息表中完备信息比例低于50%时，就可以跳过此环节直接进入决策规则的提取。

3 实验

实验一 实验一采用Lenses数据集[9]，该数据集含有24个样本，共有4个条件属性和1个决策属性，其值域分别为3,2,2,2,3，没有缺失数据。实验中，随机地选取12个样本作为训练集，并通过随机数发生器，人为地产生一定比例的缺失数据，从而得到不完备信息系统，对该系统进行决策规则提取后在原来完备的数据集中进行测试。

测试时，对于同一个样本，如果根据不同的规则能推出不同的结果，或者依据现有决策规则不能判断该样本的决策属性，均判定为识别错误。本文选用根据该判定标准在数据补齐法中识别性能最好的特定值法和基于决策的最大概率法作为对比方法。实验中，缺失数据的比例分别取25%和50%，实验次数各为3次，实验结果见表2。

表2 测试结果

缺失比例	25%			50%		
平均识别率	概率法	特定值法	本文方法	概率法	特定值法	本文方法
	68.1%	69.4%	72.2%	45.8%	45.8%	47.2%

实验二 比较特定值法、基于决策的最大概率法及本文方法在Echocardiogram、Hepatitis和House等3个不完备数据集上的识别率。

实验采用随机方式抽取50%的样本作为训练集，全数据集作为测试集，评价标准与实验一相同，实验次数各为3次，表3是测试结果。

表3 几种方法在其它不完备数据集上的测试结果

平均识别率 / 数据集	概率法	特定值法	本文方法
Echocargiogram	73.4%	75.1%	74.6%
Hepatitis	74.3%	69.8%	78.7%
House	90.5%	87.2%	92.3%

影响不完备数据集识别率的主要因素是训练集所能提供的数据关系及内在规律是否足够，以House数据集为例，虽然将近50%的样本含有缺失值，但由于各数据之间的内在关系较为紧密，因此识别率要高于同样有将近50%的样本有缺失值的Hepatitis数据集，而Echocargiogram数据集的数据内在规律相对较少，即使含有缺失值的样本所占比例并不多，但识别率并不是很高。

小结 本文针对不完备信息系统，在分析了现有的数据

补齐法和扩充法的优劣后，认为扩充模型善于寻找知识之间的内在联系，而数据补齐可以提供知识属性的表层分析，二者的结合能更好地挖掘知识的潜在规律，提高决策规则的正确性和可靠性，因此提出了基于差异关系和数据部分补齐的处理方法。差异关系相对于相似关系，当用相似性无法明确两个对象的等价关系时，我们可以转而利用它们的差异性。本文定义了差异关系，对差异矩阵进行扩充使其能适用于不完备信息系统，证明了用差异关系进行属性约简和求核的可行性，并给出了相应的算法。数据部分补齐的目的是通过分析对象之间的差异性更进一步地挖掘信息之间的潜在联系，为后续的决策规则的求取提供更丰富更准确的信息。实验证明，在处理不完备信息系统时，基于差异关系和数据部分补齐的方法能获得更好的分类性能。

参 考 文 献

1 Grzymala-Busse J W, Fu M. A Comparison of several approaches to missing attribute values in data mining. In: Proceedings of the 2nd International Conference on Rough Sets and Current Trends in Computing, Berlin: Springer-Verlag, 2000. 378～385

2 傅媛. 一种基于决策的数据补全方法. 昆明理工大学学报(理工版), 2003, 28(6): 157～160

3 张士林, 毛海军, 赵秋艳. 对于粗糙集应用中数据不全的解决方法研究. 计算机工程与应用, 2003. 10～11, 50

4 Wong K C, Chiu K Y. Synthesizing statistical knowledge for incomplete mixed-mode data. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1987, 9: 796～805

5 Kryszkiewicz M. Rough set approach to incomplete information system. Information Sciences, 1998, 112: 39～49

6 Stefanowski J, Tsoukias A. On the extension of rough sets under incomplete information. In: Proceedings of The Seventh International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing (RSFDGrC'99), 1999. 73～81

7 朱小飞, 卓丽霞. 一种基于量化容差关系的不完备数据分析方法. 重庆工学院学报, 2005, 19(5): 23～25

8 余英泽, 王国胤, 吴渝. 一种基于 Rough 理论的不完备信息系统处理方法. 计算机科学, 2001, 28(5): 35～38

9 Witten I H, MacDonald B A. Using concept learning for knowledge acquisition. International Journal of Man-Machine Studies, 1988, 27: 349～370

(上接第 148 页)

这组实验(图 4、图 5)主要考察结果集大小对查询时间的影响。随着查询结果集增大，LBD、VA-file 和 iDistance 查询时间缓慢增加，但是总体上 LBD 查询时间明显少于 VA-file 和 iDistance 所用时间，只有 1/6 左右。LBD 索引利用分区，可以将数据合理地划分到不同的分区中，便于检索时剪枝，使得 LBD 更加适用于非均匀分布数据。

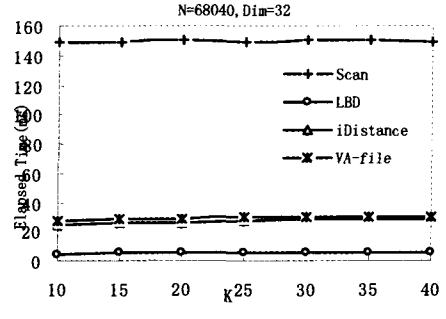

图 4 查询范围对查询时间的影响

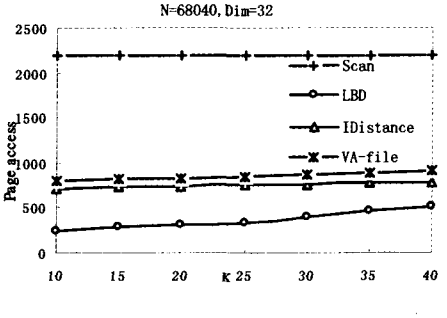

图 5 查询范围对页面访问次数的影响

综合可知，LBD 的查询效率远远优于 SCAN、iDistance 和 VA-file，尽管它们都是高维数据空间检索表现较好的索引结构，特别是随着数据量增大或维度增高这种优势更加明显，并且 LBD 索引结构不受数据分布影响，能很好地适用于各种数据分布情况，所以 LBD 是一种有效的针对高维空间 KNN 检索的索引结构。

结论 本文提出了一种有效的高维数据表示(LBD)，基于此给出一个简单有效的 KNN 搜索算法。在充分考虑到高维空间向量分布特点的前提下，LBD 对高维空间进行聚类划分利用距离索引组织 B^+ 树，并对每个聚类子空间合理分区，通过简单的部分维上的位码字符比较，大大提高 KNN 搜索剪枝效率。实验证明，LBD 的检索性能优于其他同类 KNN 搜索算法。

参 考 文 献

1 Chavez E, Navarro G, Baeza-Yates R, et al. Searching in metric spaces. ACM Computing Surveys, 2001, 33(3): 273～321

2 Fonseca M J, Jorge J A. Indexing high-dimensional data for content-based retrieval in large databases. In: Proceedings of the Eighth International Conference on Database Systems for Advanced Applications (DASFAA'03), 2003. 267～274

3 Hjaltason G R, Hanan S. Index-driven similarity search in metric spaces. ACM Transactions on Database Systems, 2003, 28(4): 517～580

4 Weber R, Schek H, Blott S. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In: Proc 24th International Conference on Very Large Data Bases, 1998. 194～205

5 Beyer K, Goldstein J, Ramakhrisnan R, et al. when is "Nearest neighbor" meaningful. In: Proc. the 7th Int'l Conf. Database Theory (ICDT'99), 1999. 1～11

6 Hinneburg A, Aggarwal C C, Keim D A. What is the nearest neighbor in high dimensional spaces. In: Proc. 26th International Conference on Very Large Data Bases, 2000. 506～515

7 BinCui, et al. Exploring Bit -Difference for Approximate KNN Search in High-dimensional Databases. In: Proc. 16th Australasian Database Conference, 2005. 165～174

8 Berchtold S, Bohm C, Kriegel H P. The pyramid-technique: Towards breaking the curse of dimensionality. In: Proc. ACM SIGMOD International Conference on Management of Data, 1998. 142～153

9 Yu Cui, et al. Indexing the Distance: An Efficient Method to KNN Processing. In: Proc. 27th International Conference on Very Large Data Bases, 2001. 421～430

10 Chakrabarti K, Mehrotra S. Local Dimensionality Reduction: A New Approach to Indexing High Dimensional Spaces. In: Proc. Conf Very Large Data Bases, 2000. 89～100

11 Jin H, Ooi B, Shen H, et al. An Adaptive and Efficient Dimensionality Reduction Algorithm for High-Dimensional Indexing. In: Proc. Int'l Conf. Data Eng, 2003. 87～98

12 Corel Image Features. http://kdd.ics.uci.edu