使用背景知识处理不完全信息*)

Dealing with Uncertainty Information Based on Background Knowledge

王拥军　何华灿

(西北工业大学计算机科学与工程系　西安710072)

Abstract　In past years, people were hardly aware of Mapping Knowledge between Environments when dealing with problems, did not consider the ability of their constraint and introduction to handle cognition problem under incomplete information environment. This paper firstly points out the importance of explicit representation on background knowledge, then gives a method that deals with imprecise information aggregation using generalized rough set theory. This process involves representation and acquiring about background knowledge.

Keywords　Generalized rough sets, Mapping knowledge between environments, Aggregation, Background knowledge

1　引言

不完全性信息大致可以分为两类：其一表现出不精确性，即属性值的内容不唯一[1]；其二呈现不确定性，即属性值的真值程度不肯定。这两种情形下许多问题的通用求解往往是NP问题，实际中只能在特定的约束和限制下解决。

尽管人们很少意识到环境映射知识的存在，然而它却在人类认知和智能活动中扮演了十分重要的角色。人们在遭遇一个相对陌生的环境时，总是试图找出一个与之相类似的熟悉场景作参考，继而获得一个加之于原问题之上的弱结构关系(不一定完全正确)，使不完全性信息下的问题求解摆脱搜索空间过大、知识表达不清晰等的困扰。

本文给出一种通过相似环境映射获得欲处理问题背景知识的方法，并用结构关系R显式表示。具体用不精确信息的处理过程例示说明之。

2　背景知识的获取方式

当我们做关于经验世界的判断和推理时，总会涉及到不同环境之间一些种类的映射。这类映射与事实世界无关，着重于可能环境间的关系，能得到不一定正确但有用的新环境背景知识。这样的背景知识随着新信息的加入有可能会发生调整，因而是非单调的，可以适应由于信息不完全和环境动态变化引起的约束改变。

环境映射知识就是管理上述映射的知识。根据客观世界中广泛存在的相似性、可以将以前曾经遇到过的环境中的一些相关特征转移到新环境中[2]。同时，由于没有绝对的重复，不同环境间又必然有差异。如何正确地辨别出哪些特征对新环境有用、哪些可以忽略，是环境映射知识最主要的任务。

本文用结构关系R来表示获得的特定背景知识。R可以是概念层次结构，可以是一个二元关系，也可以是专家定义的其它结构形式。这主要取决于环境映射知识起作用的方式，以下给出三种环境映射的方法。

·首先是归纳推理形式[3]。人类认知时通过不断累积、归纳以形成对一类事物的认识，可以作为通用知识指导本类新生事物的再认知。由于归纳过程中不能覆盖此类事物的全体，故而向新事物的推广必然面临如何根据具体环境选择通用知识的投影方式问题。这种映射方法的关键在于新事物与通用形式的差异，现时一般采用与讨论问题的相关程度作为属性取舍的依据。

·其次是类比推理形式[4]。与归纳推理不同，这是一种横向的推理方式。它从一个典型事例抽取与之相似情形的一般特征，然后将新环境作为此通用状态的示例来填充。这个通用状态不是显式给出的，而是为类比的结果所隐含，不难看出，此种推理的困难在于找出合适的一般化抽取，为了执行合理的类比推理需要对

*)本文得到国家教委博士学科点专项科学基金(98069923)和陕西省自然科学基金(98X15)的资助。王拥军　博士生，主要从事人工智能基础理论、不确定性推理和需求工程的研究。何华灿　博士生导师，主要研究人工智能基础理论和泛逻辑等。

环境不同方面的取舍有一个良好的判断。此外，从认识新环境的不同层次和角度出发，选择的典型环境也不尽然相同，怎样得到当前知识下最合适的类比原型也是一个棘手的问题。

·最后是反绎推理形式，其主要思想是找出一个具体环境的成因[5]，再用它来解释此环境中各种现象。难点在于辨别对潜在原因有影响的方面和仅是表面一致的方面，即如何抓问题的主要方面，使得生成的原因有更广泛的适用范围。

综上所述，在推理中起关键作用的是环境映射知识，它决定着推理过程中如何正确地从一般状态向具体环境推广，如何选择具体环境的特征属性，如何处理几种推广之间的矛盾等关键技术的实施。基于同样经验数据的专家通常比外行可以进行更好的归纳推理，这表明专家执行的"有目的"猜想要远远优于外行进行的"公平"猜想。专家的这种能力源于其对领域的熟悉，知道何种属性和客体与具体推广有更大的相关性，便于为新环境产生一个更有效的背景。

本文建议使用给欲讨论的环境属性赋予不同的优先级，赋值原则是：在一个具体环境中属性值越容易改变，则此属性对该环境的重要性越低。可以知道，不同环境中同一属性的相对重要程度是不同的，依赖于具体的上下文，在形成通用状态时应当综合考虑、协调处理。

当然，还有一些不成体系但又经常有效的专家知识，它能对上述三种方法形成的结构关系 R 做修正和补充，使得应用于具体环境的背景知识更具指导和约束能力。

3 背景知识对问题的约束作用

环境的背景知识一旦形成，就相当于给出了一种具体的假设，是环境映射知识关于可能世界集合的一个当前最合理取值。用结构关系 R 表示的背景知识，可以用加于原问题结构之上的覆盖图描述之。若此图是一棵树，则 R 就是一个层次结构。

在带有结构约束的模糊聚类问题中，R 是定义在数据样本集合之上的一个二元关系，由对类间关系的了解给出。这种了解往往是根据归纳、类比和专家知识得到的，有很强的针对性和分类指导作用。同时，R 又带有一定的猜测成分，因而聚类问题亦具有随环境改变做相应调整的动态性质。

其它的 R 结构约束问题，虽然 R 的形式不尽相同，但大致处理过程相近。下面通过对一个具体的不精确信息系统的处理，给出本文对这类问题的解决思想。

4 不精确性信息的处理

在数据库中由于诸种干扰，经常会出现数据丢失或不合规范等情形；有时也会因为对客观事物了解的不完全，使得属性值出现不知道情形。这都会导致空值现象[6]，使得数据处理无法正常进行。本部分就是基于环境间广泛存在的相似关系，利用环境映射知识给出当前讨论环境的背景知识，然后在此约束下重新整理数据库，最后给出如何处理新形式数据的方法。

4.1 带有空值的数据库

现实生活中，我们会接触到各种各样的新环境，且经常是在无足够导引的情况下进行认知过程。信息的不完全性使得数据的属性值出现空值情形，即不知道其应为何值。表1给出一个甲大学学校职工的关系数据库。

表1 甲大学学校职工关系库

姓 名	性别	工作岗位	岗位工资(元)
赵云飞	男	科长	空值
钱文长	男	职工	260
孙成都	男	系主任	682
李道明	女	教师	400
周 峰	男	空值	424
吴满有	男	校长	776
郑平顺	男	空值	676
王若非	男	处长	654
冯 丽	女	一般干部	390
陈李开	男	空值	723

从表中可以知道，当我们不熟悉属性有空值的客体时，数据查询和处理就都无法进行。背景知识的引入增强了对这些所谓"不知道"情形的了解，使得空值变成了部分可知的"可能值"，可以得到在当前环境下令人满意的结果。

4.2 根据环境映射知识形成具体的背景知识

即使不很熟悉正处理的数据库所对应的环境，我们仍然可以通过环境间的相似性映射获得更多新环境的背景知识，这些经验知识被证明对于迅速把握问题的关键是行之有效的。

环境映射知识能指导我们从各种各样的已识别环境中归纳出大量有用的通用知识，也能够找出最相似的一个环境作为参照来提供新环境所需的背景知识。从对上表中存在空值的属性"工作岗位"和"岗位工资"两项分析可知，其工资的高低反映了岗位的不同重要程度、即岗位之间的关系是与此问题紧密相关的背景知识。

由于对普遍存在的管理体系和一些其它学校的了解，不难给出一个基于层次关系的结构关系 R，见下面图1。

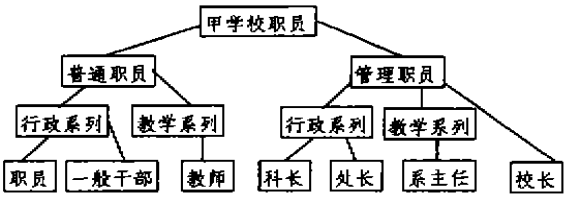

图1 岗位关系概念层次结构

从上图可以知道的背景知识有：管理职员的岗位比普通职员的岗位重要；处长岗位优于科长岗位；校长是甲学校最重要的岗位；一般干部岗位先于职员岗位。这些都是经过一般管理常识向新环境的推广，仅从它们不能得出教学系列和行政系列间的岗位关系。

此时系统会根据以往所遇到的有关学校的数据库相关情形，给出教学系列和行政系列间的岗位关系。若有几个此类数据库可供选择，则应选出和甲学校最为相近的学校作为参考。若一个也没有，就寻找有无专家知识作指导，这都有助于结构关系 R 的形成。本文假定有一中学数据库可供参照，由此可得出教师和一般干部岗位同等重要的结论。另外，从专家知识"系主任是正处级干部"可假设系主任和处长岗位同等重要；且同一岗位的工资相差在10%内是正常的。

以上就是利用环境映射知识得出的新环境甲学校的背景知识（尽管实际上可能不尽如此），是一种弱结构关系，可以显式给出。

4.3 基于背景知识生成广义数据库

有了层次结构关系 R，就可以重整含有空值的数据库了。重整的含义是根据显式的背景知识将原来"不知道"的空值用"可能值"代替，这个"可能值"是指该论域中必定包含真值的子集。重整后的数据库称为广义数据库。

假定工作岗位属性提供的最下层概念论域是{职员，一般干部，教师，科长，处长，系主任，校长}。关系库中赵云飞是一名科长，故由背景知识可知他的岗位工资应当高于普通职员的最高值，同时低于处长或系主任这一级的最低值。很快可以得出工资区间是[400，624]，注意这里实际上是离散取值。这里还可以根据10%的正常误差再次给出更好的估计[440，562]，且一般来说较前面的估计更合理。周峰的岗位工资低于管理职员中岗位最低的科长工资下限，所以不可能是管理人员。用10%正常误差来衡量，发现他的工资处于一般干部和教师两种情形的合理范围内，所以可能取值集合为{一般干部，教师}。郑平顺的岗位工资正好位于同级的处长和系主任之间，且都是他们的正常范围之内，所以取值{处长，系主任}。陈季开的岗位工资能落在处长、系主任和校长三种选择的合理范围之内，但由常识"校长的岗位最重要"可知其不可能是校长，因此取值{处长，系主任}。表2是用可能值代换空值的广义数据库。

表2 甲大学学校广义职工关系库

姓 名	性别	工作岗位	岗位工资(元)
赵云飞	男	科长	[440，562]
钱文长	男	职工	260
孙成都	男	系主任	682
李道明	女	教师	400
周 峰	男	{一般干部，教师}	424
吴满有	男	校长	776
郑平顺	男	{处长，系主任}	676
王若非	男	处长	654
冯 丽	女	一般干部	390
陈季开	男	{处长，系主任}	723

4.4 利用广义粗集理论度量不精确信息

广义数据库中存在"可能值"，故需用不确定性度量方法实现信息的聚合。广义粗集理论[7,8]提供了度量和处理不精确数据的简单方法、据此可以形成聚合算子，用广义上、下近似集分别给出"可能值"的界限。

相容关系的给出。由于广义数据库中属性值定义与以往不同、划分有可能出现相互覆盖的情形。这里将等价关系中的传递性去掉("可能值"的引入使之失效)，成为相容关系。上例按工作岗位划分，则一个覆盖为：{{赵云飞}$_{科长}$，{钱文长}$_{职员}$，{孙成都，郑平顺，陈季开}$_{系主任}$，{王若非，郑平顺，陈季开}$_{处长}$，{冯丽，周峰}$_{一般干部}$，{李道明，周峰}$_{教师}$，{吴满有}$_{校长}$}。

可能集合的度量。广义上、下近似的定义基本上与基于等价关系的上、下近似定义相同，只不过是用覆盖代替等价类而已。工作岗位属性上所有值认为出现几率相当且相加为1。

聚合算子的使用。现在考虑带有可能值的数据模型聚合问题，它的实现可以提供相关属性每种可能取值的汇总情况。假设 A 为关系数据库内一属性，论域 $D=\{V_1,\cdots,V_k\}$，元组 $T=\{t_1,\cdots,t_n\}$，t_i 关于 A 属性的值 $d_i\subseteq D$。当给定一种可能取值时，先给出它的广义上、下近似，然后分别求出两个近似集合的出现几率和，构成的区间作为对可能值的度量。

定义1 n 个元组使用聚合算子 agg 进行聚合，其中 $agg(A)=\{(d,R_d)\}$，这里 d 是可能值(精确值是特殊的可能值)，R_d 为 d 取值的几率和，$R_d=(1/n)\sum_{i=1}^{n} I(d=d_i)$，I 是标志函数，当条件满足时，取1，否则是0。

上表中 R({科长})=R({处长})=R({教师})=R({职员})=R({校长})=R({一般干部})=R({一般干

(下转第58页)

测试情况	示例文本数	测试文本		准确率	查全率
		相关文本	不相关文本		
测试一	8	10	3	78%	86%
测试二	10	8	6	83%	85%
测试三	20	15	15	81%	88%
测试四	30	10	20	92%	85%
测试五	30	50	30	86%	89%
测试六	10	50	50	82%	85%
测试七	50	30	10	91%	93%
测试八	50	50	50	90%	91%

结束语 随着 Internet 的迅速普及，WWW 上的信息资源越来越丰富，更多的用户要求从网上获取信息。但是网络信息的快速膨胀，使得人们越来越难以获得符合自己要求的数据信息。因此，广大网民迫切需要一个具有自学习功能的信息过滤器，提高信息获取的效率和信息的质量。

在文章中，我们提出了一个基于示例学习的文本过滤器模型。在该模型中，我们采用改进遗传算法来获取 Web 文本的特征向量，从而有效地解决了传统文本处理过程中，文本向量维数过大的问题，提高了系统处理的效率。另外，我们基于语义的模糊特性，采用了模糊技术来表示特征向量，进一步提高了过滤模型的处理精度。从实验结果可以看出，该模型基本满足文本过滤处理的要求，具有令人满意的精度。

参 考 文 献

1 Gudivada V N. Information retrieval on the World Wide Web. IEEE Internet Computing, 1997, 1(5): 58～68

2 Mladenic D. Machine Learning on non-homogeneous, distributed text data, doctoral dissertation. University of Ljubjana, 1998

3 Mladenic D, Grobelnik M. Feature selection for classification based on text hieraracy. Working Notes of learning from Text and the Web. Conf. on Automated learning and Discovery CONALD-98, 1998

4 张晓辉. WWW 上的信息发现与搜索引擎技术. 小型微型计算机系统, 1998, 19(6): 66～71

5 王实, 高文, 李锦涛. Web 数据挖掘. 计算机科学, 2000, 27(4), 28～31

6 张月杰, 姚天顺. 基于特征相关性的汉语文本自动分类模型的研究. 小型微型计算机系统, 1998, 19(8): 49～55

7 王伟强, 高文, 段立娟. Internet 上的文本数据挖掘. 计算机科学. 2000, 27(4): 32～36

8 吴立德, 等. 大规模中文文本处理. 复旦大学出版社, 1997

9 王继成, 潘金贵, 张福炎. Web 文本挖掘技术研究. 计算机研究与发展, 2000, 37(5), 513～520

10 邹涛, 王继成, 等. WWW 上的信息挖掘技术及实现. 计算机研究与发展, 2000, 36(8): 1019～1024

11 陈恩红. 基于遗传算法的机器学习方法及应用研究: [博士论文]. 中国科学技术大学, 1995

12 Liu Mingji, Wang Xiufeng. A Knowledge Discovery Algorithm Based on Genetic Algorithm. The Third World Congress on Intelligent Control and Automation, IEEE WCICA'2000

13 Victor Ciesielski, Zhnag Memgjie. Using Genetic Algorithms to Improve the Accuracy of Object Detection. PAKDD-99, 19～24

14 Rudolph G. Convergence Properties of Canonical Genetic Algorithms. IEEE Trans on Neural Networks, 1994, 5(1): 96～101

(上接第39页)

部，教师}) = R({系主任}) = 0.1, R(处长，系主任) = 0.2。

对任意取值{处长}，其广义上、下近似分别为{处长}和{{处长}∪{处长，系主任}}，故度量区间是(0.1, 0.3)。其余可同理推出，不再赘述。

对于大型数据库，作用更加明显。

结论 用环境映射知识形成背景知识的过程本身又是一个系统学习的过程。通过对新环境的假设和对不同假定之间矛盾的处理，系统会提高归纳、类比等抽取知识的能力，并在运行过程中调整已有的关于环境间映射的知识，这对如何把握其它相似环境乃至更复杂环境提供了认知基础。

客观世界本质上的非全息性使得表示和处理都必须面对不确定性问题，本文给出了一种如何解决数据库中存在不精确信息问题的方法。

参 考 文 献

1 Dubois D, Prade H. Possibility Theory: An Approach to Computerized Processing. New York: Plenum Press, 1988

2 Blum A L, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence, 1997, 97: 245～271

3 洪家荣. 归纳学习. 科学出版社, 1997

4 Medin D L, Goldstone R L, Gentner D. Respects for Similarity. Psychol. Rev., 1993, 100: 254～278

5 孙吉贵, 刘瑞胜, 陕荣. 不完全信息下的溯因诊断. 吉林大学自然科学学报, : '98, 10(4): 34～38

6 Chen A L P, Tseng F S C. Evaluating aggregate operations over imprecise data. IEEE Transactions on Knowledge and Data Engineering, 1996, 8: 273～284

7 Slowinski R, Vanderpooten D. A Generalized Definition of Rough Approximations Based on Similarity. IEEE Transactions on Knowledge and Data Engineering, 2000 12: 331～336

8 Bonikowski Z, et al. Extensions and intentions in the rough set theory. Information Sciences, 1998, 107: 149～167