知识发现状态空间模型研究及其应用*)

游福成[1,2]　杨炳儒[2]

(北京印刷学院计算机系　北京102600)[1]　(北京科技大学信息工程学院　北京100083)[2]

摘　要　结构化数据挖掘与复杂类型数据挖掘既有联系，又有区别。如何将这两者统一起来，建立一个统一的理论框架，以指导数据挖掘与知识发现研究，已经成为一个迫切需要解决的问题。本文提出了知识发现状态空间统一模型UMKDSS，将结构化数据挖掘与复杂类型数据挖掘联系起来，为复杂类型数据挖掘提供理论指导。文章最后给出了UMKDSS在Web文本挖掘中的应用实例。

关键词　知识发现，非结构化数据，知识模板

The Research of United Model of Knowledge Discovery State Space and Its Application

YOU Fu-Cheng[1,2]　YANG Bing-Ru[2]

(Department of Computer, Beijing Institute of Graphic Communication, Beijing 102600)[1]

(Information Engineering School, University of Science and Technology, Beijing100083)[2]

Abstract　There are both associations and differences between structured and unstructured data mining. How to unite them together to be a united theoretical framework and to guide the research of knowledge discovery and data mining has become an urgent problem to be solved. On the base of analysis and study of existing research results, this paper puts forward the United Model of Knowledge Discovery State Space (UMKDSS), and associates the structured data mining and the complex type data mining together. UMKDSS can provide theoretical guidance for complex type data mining. An application example of UMKDSS is given in the end of this paper.

Keywords　Knowledge discovery, Unstructured data, Knowledge templet

1　概述

目前数据挖掘对象已经扩展到复杂类型数据，如多媒体、音频、视频、图形图像、Web内容、空间数据、时序数据等。结构化数据挖掘与复杂类型数据挖掘既有联系，又有区别。如何将这两者统一起来，建立一个统一的理论框架，以指导数据挖掘与知识发现研究，已经成为一个迫切需要解决的问题。

有一些研究者从不同的角度提出了不同的理论框架，如证据理论(Evidence Theory)，Rough集理论等。李德毅提出了发现状态空间理论[1]，邸凯昌博士把发现状态空间理论进行拓展，形成了空间知识发现状态空间理论SDMKD[2]。两者的区别在于后者增加一个空间尺度维，使扩展后的发现状态空间变成一个四维发现状态空间。

但两者都不适合非结构化数据和复杂类型数据挖掘，因为对于诸如图像图形数据、文本数据、时序数据等复杂数据的挖掘过程中进行的特征空间降维、特征提取和特征变换等特点，它无法描述和刻画出来。

在分析与研究前人研究成果的基础上，本文提出了知识发现状态空间统一模型UMKDSS(United Model of Knowledge Discovery State Space)，目的就是要将结构化数据挖掘与复杂类型数据挖掘联系起来，形成一个统一整体。

2　UMKDSS的结构模型研究

通过研究结构化数据与复杂类型数据的特点和内在规律，结合知识发现的本质规律性，在分析知识发现研究现状的基础上，笔者提出了知识发现状态空间统一模型UMKDSS，如图1所示。

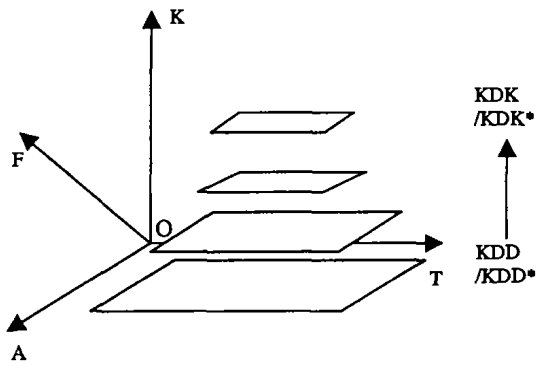

图1　知识发现状态空间统一模型UMKDSS

从整体来看，知识发现状态空间统一模型UMKDSS是一个四维空间，其中OA轴为属性(Attribute)轴，代表研究对象的属性或字段；OT轴为元组(Tuple)轴，代表研究对象的各个属性集合的元组或记录；OK轴为知识模板(Knowledge Templet)轴，代表知识的抽象层次；OF轴为特征(Feature)轴，代表复杂类型数据的特征。

从结构来看，OA轴、OT轴和OK轴组成的空间是结构化数据挖掘的运作空间，而OF轴则是复杂类型数据向结构化数据近似转换的方向。因此，知识发现状态空间统一模型

*)本课师得到了国家自然科学基金重点项目(69835001)及教育部科技重点项目(教技司[2000]175)的支持。杨炳儒　教授，博士生导师，研究领域为数据挖掘与知识发现、柔性建模与集成技术；游福成　博士研究生，研究领域为知识发现。

UMKDSS 把结构化数据挖掘与复杂类型数据挖掘有机地结合在一起，形成了知识发现状态空间的统一理论框架。

在知识发现状态空间进行的多种知识汇集和发现操作分成四个方向，即面向属性的操作、面向元组的操作、面向知识模板的操作和面向特征的操作。人工神经网络方法、证据理论、粗集理论、统计理论、概念格方法、概念树提升方法、遗传算法、因果关系定性推理方法、演绎与归纳方法等知识发现方法都分别直接或间接地要涉及到这四个方向的操作；要发现诸如分类知识、聚类知识、相似模式知识、关联规则等也要涉及这四个方向的操作。

2.1 OA 轴方向——面向属性的操作

在 OA 方向的操作是面向属性的操作，是对属性之间关系的认识和发现活动，主要有规格化、剪枝、并枝等。其中规格化是指属性值量纲的归一化处理；剪枝是去掉对知识发现任务没有贡献或贡献率极低的属性域；并枝是将相近属性进行综合、归并处理，如主成分分析、属性降维等技术。

大量的数据库或数据库中的部分属性是随时间变化的，时态数据库技术研究的就是随时间而变化的数据库的存储、管理和操作。在知识发现状态空间中，我们把时间看成是一个特殊的属性，针对时间的操作仍然是面向属性的操作，比如，时间序列分析研究的是其他属性与时间属性的关系。

2.2 OT 轴方向——面向元组的操作

在 OT 方向的操作是面向元组的操作，是对各种元组之间一致性和差异性的认识和发现活动。相应的操作有排序、分类和聚类操作。排序是指在一定的条件约束下，使元组形成一种规则排列；分类是已经知道确切的分类数和每类的典型特征，对所有元组进行归类的过程，属于有导师的学习过程；聚类是事先不知道要分成多少类，也不知道各类的典型特征，完全由元组属性实际分布的聚散程度来决定类的划分，使得类间相似性最小，类内相似性最大，属于无导师的学习过程。

2.3 OK 轴方向——面向知识模板的操作

在 OK 轴方向的操作是面向知识模板的操作，是从微观到宏观的发现知识的操作。主要操作有概念树的生成和调节、规则置信度阈值的调节、域间抽象层次适配性检查、概念提升、发现知识的验证和评价等。

此外，OK 轴方向有 KDD/ KDD*、KDK(Knowledge Discovery in Database) /KDK* 两种知识发现的抽象过程，从而可以把我们的研究成果——由 KDD* 和 KDK* 构成的知识发现内在机理 KDTIM 有机地容纳进来。

2.3.1 知识发现 KDD　KDD 模型如图2所示，它包括以下处理步骤：

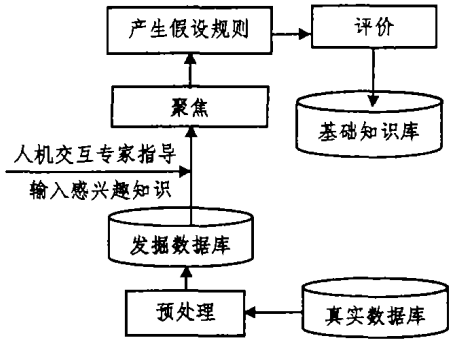

图2　KDD 模型

(1)数据选择：根据用户的要求从数据库中提取相关的数据，形成真实数据库。

(2)数据预处理：主要是对步骤(1)产生的数据进行再加工，检查数据的完整性及数据的一致性，形成挖掘数据库。

(3)确定 KDD 的目标：根据用户的要求，确定 KDD 是发现何种类型的知识。

(4)确定知识发现算法：选择合适的知识发现算法，包括选取合适的模型和参数，并使得知识发现算法与整个 KDD 的评判标准相一致。

(5)聚焦：即从挖掘数据库里进行数据的选择。指导数据聚焦的方式是通过人机交互由专家输入感兴趣的知识，来指导数据的挖掘方向。

(6)产生假设规则：运用选定的知识发现算法，从数据中提取出用户所需要的知识，这些知识可以用一种特定的方式表示或使用一些常用的表示方式，如产生式规则等等。

(7)知识评价：这一过程主要用于对所获得的规则进行价值评定以决定所得的规则是否存入基础知识库。主要是通过人机交互界面由专家依靠经验来评价。

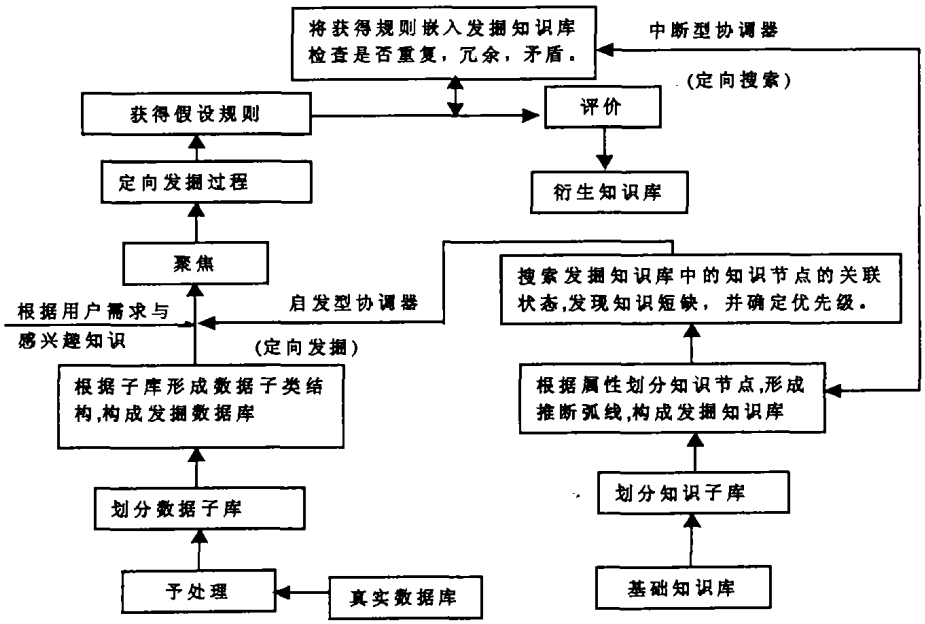

图3　KDD* 系统总体结构图

2.3.2 基于双库协同机制的知识发现KDD* KDD*是指基于双库协同机制的知识发现[3,4],与KDD相比,KDD*加入了双库协同机制,构建了两个协调器:启发型协调器和维护型协调器,是KDD的一个优化扩体。在结构对应定理的基础上,它通过启发型协调器从知识库中发现知识短缺,定向启动挖掘进程使系统产生自动聚焦,得到假设规则;并通过维护型协调器实时地到知识库中对应位置进行定向搜索,以查找是否存在重复、冗余、矛盾的规则,进行知识库的实时维护。其结构如图3所示。

2.3.3 知识库中的知识发现KDK KDK是指知识库中的知识发现,KDK的描述性界定为:

(1)KDK的目的是为了在真实的大型知识库中发现新的知识。这种发现过程的核心将是归纳,而演绎将作为辅助手段,该过程不同于传统的演绎,它有可能是不保真的。

(2)KDK能够发现深层次的知识。具体而言就是在已有关系的基础上进一步发现其上的关系,从逻辑角度上说就是发现谓词间的关系或涵词间的关系。

(3)由于知识本身所可能具有的一些属性,如不确定性,非单调性,不完全性等,KDK过程也将是一个涉及多方法多途径的过程。它与知识库的组织,用户对最终寻求的知识的类型都紧密相关,采用的推理手段可能涉及很多不同的逻辑领域。

(4)KDK发现的知识应该是有效的,潜在有用的,用户可理解的。

从上述的定义我们可以看出,KDK究其本质来说应该是一种机器学习过程,其本质目的是获取知识,学习源是知识库,学习手段是用归纳结合演绎的方法,其最终结果将既能够发现事实上的知识,也能发现关系上的知识。

2.3.4 基于双基融合机制的知识库中的知识发现KDK* KDK*是基于双基融合机制的知识库中的知识发现,其过程模型如图4所示。所谓双基融合机制是指构建基础数据库与知识库的内在联系的"通道",从而用基础数据库去制约与驱动KDK的发掘过程,改变KDK固有的运行机制,在结构与功能上形成了相对于KDK而言的一个开放的优化的扩体,包含以下四个方面含义:

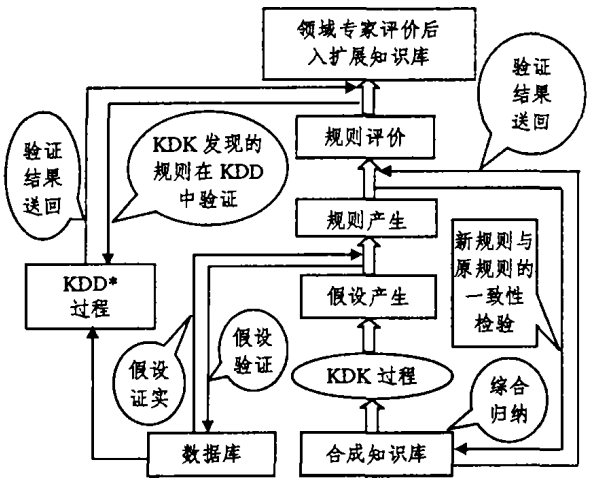

图4 基于双基融合机制的面向规则的KDK*示意图

(1)KDD过程依赖于知识库。即在KDD的发掘过程中要随时与知识库中的现有知识相关联,以先验知识来约束和推动KDD过程。

(2)KDK过程要依赖于数据库。即将在KDK过程中发掘出的新知识随时送入数据库中,以数据库中的数据来验证新知识。

(3)KDK的过程要依赖于KDD的发掘过程。即在KDK中发掘出的新知识要随时送入KDD过程中进行验证,看KDK中发现的规则在KDD过程中是否能被发现。这一过程本质上是确定KDK中发现的新规则在现实中是否有意义。

(4)KDK依赖于原有知识库。即KDK过程中发现的新知识要实时地带入原有知识库,以验证新知识与原有知识是否冗余、矛盾和重复。

从以上揭示的四种内涵我们可以看出,双基融合机制的本质在于数据库与知识库的一种协调,这种协调不是简单的叠加,而是挖掘出两者的内在联系,从本质上寻求切入点。将数据库和知识库统一在一个系统中,使它们能够相辅相成,是一种机器智能的较高境界。

知识模板可以看成是人们对现实世界中特定客体在一定抽象级别上的观察记录和认识,模板结构反映出隐含的知识量和发现知识的复杂程度。我们定义知识熵H来度量知识量,发现难度系数D来度量发现知识的难度。其中知识熵表示知识模板中所蕴涵的知识不确定性的状态。

定义1 设知识模板K含n个簇,其中第s个簇覆盖的元组数为W_s,则整个知识模板覆盖的总元组数为:

$$N=\sum_{s=1}^{n}W_s$$

又设从该模板中可能得到的知识(规则)条数为R条,支持第$i(i=1,2,\cdots,R)$条规则的簇数为M_i,其中第M_j个簇覆盖元组数为T_{ij},则第i条规则覆盖的元组数为

$$N_i=\sum_{j=1}^{M_i}T_{ij}$$

定义知识模板K对应的知识熵为H:

$$H=-\sum_{i=1}^{R}\frac{N_i}{N}*\log_2\frac{N_i}{N}$$

定义2 设知识模板K中含有λ个属性:$(A_1,A_2,\cdots,A_j,\cdots,A_\lambda)$,其中第$j$个属性$A_j$有$\beta$个取值$\{a_j^1,a_j^2,\cdots,a_j^\beta\}$,$j=1,2,\cdots,\beta$;若按照$A_j$值来划分知识模板$K$,可得到$\beta$个知识子模板$\{K_j^1,K_j^2,\cdots,K_j^\beta\}$,其中$K_j^q$含有$n_j^q$个簇,这些簇在$A_j$这个属性中具有相同的$a_j^q$值。

设知识子模板K_j^q的知识熵为H_j^q,则对属性A_j来说,知识熵为:

$$H_j=\sum_{q=1}^{\beta}\frac{n_j^q}{N}*H_j^q$$

因此,由属性A_j引起的知识增益为:

$$G(A_j)=H-H_j$$

因为属性A_j表现出来的信息量为:

$$I(A_j)=-\sum_{q=1}^{\beta}\frac{n_j^q}{N}*\log_2\frac{N_j^q}{N}$$

因此,知识发现难度系数为:

$$D=\prod_{j=1}^{\lambda}[G(A_j)/I(A_j)]$$

知识熵H和知识发现难度系数D可以作为知识发现过程的目标函数,指导整个发现活动和思维机制。

2.4 OF轴方向——面向特征的操作

OF轴方向的操作是针对非结构化数据(Web数据、文本数据、图形图像数据、时序数据、空间数据等)的特征操作。主要的操作有特征空间降维、子空间逼近、特征提取、特征变换

(下转第212页)

ation-based Text Classification. In:Proc. of ACM Int. Conf. on Information and Knowledge Management, 2000
8 Bekkerman R,El-Yaniv R,Tishby N,Winter Y. Distributed Word Clusters vs. Words for Text Categorization. Journal of Machine Learning Research, 2003,3:1183～1208
9 Quinlan J R,Carneron-Jones R M. FOIL:A Midterm Report. In: Proc. European Conf. Machine Learning,1993. 3～20
10 Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. Addison-Wesley, 1999
11 Fung B C M, Wang K. Hierarchical Document Clustering Using Frequent Itemsets. In:Proc. of SIAM Intl. Conf. on Data Mining, 2003
12 Borgelt C. Efficient Implementation of Apriori and Eclat. In:Proc of the first Workshop on Frequent Itemset Mining Implementations, 2003
13 Frakes W B, Baeza-yates R. Information Retrieval: Data Structures and Algorithms, Prentice-Hall, 1992
14 Zaiane O R,Antonie M. Classifying Text Documents by Associating Terms with Text Categories. In:ADC, 2002. 215～222
15 Freitas A A. Understanding the Crucial Differences Between Classification and Discovery of Association Rules-A Position Paper. SIGKDD Explorations,2000,2(1):1～5
16 http://www.jihe.net/datasets.htm
17 The 20 Newsgroups Dataset. http://people.csail.mit.edu/u/j/jrennie/public_html/20Newsgroups/
18 李曲,冯剑琳,邹晶,冯玉才. SAT-FOIL: Sentence as Association Transaction for Text Classification. 已投稿

(上接第196页)

等。通过这些操作,非结构化数据可以近似地转化为结构化数据,然后再进行数据挖掘,这是非结构化数据挖掘的必由之路。

邸凯昌提出的四维空间知识发现状态空间也可归纳进来,也是本统一框架理论的一个特例。因为空间尺度维代表的是不同的分辨率的变化,它与图像的放大、缩小操作(使得面状目标变成点状目标),以及分辨率处理是一致的,也属于特征的一种特殊变换,因此,空间尺度维可以合并到特征维 F 轴上来。

3 应用

知识发现状态模型已经为 Web 文挖掘提供了理论指导,并取得了很好的研究效果[5]。Web 文本挖掘系统界面如下图5、图6所示,Web 文本挖掘原型系统的开发环境是 Windows 操作系统,数据库管理系统使用了微软的 SQL Server 数据库,开发工具使用了面向对象的集成开发工具 Delphi6.0。考虑到文章篇幅所限,Web 文本数据挖掘应用过程简述如下:

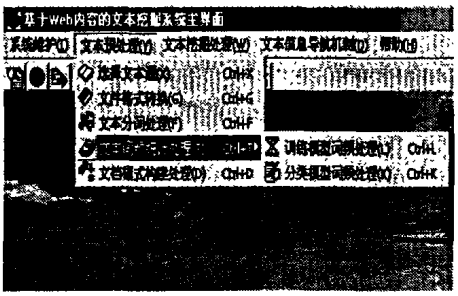

图5 Web 文本挖掘系统主界面(1)

(1)文本数据预处理过程:包括特征表示、文本信息的分词预处理、特征提取等过程。其中特征表示是指以一定的特征项(如单字或词条等)来代表文档信息;文本信息的预处理主要包括英文文档的词干处理和中文文档的词条切分;特征提取是指特征表示中词条 T 的选取,是挖掘特征共性与规则的提取过程。

(2)文本数据的知识发现算法:基于模式的文本挖掘是一个发现新模式或对模式进行某种确证的过程。通过模式矢量,可同文本分类、聚类、关联模式等收敛型的知识发现算法及预测、时序等发散型的知识发现算法相结合,来完成对于各种文本数据的知识发现。

(3)模式的评价:文本模式的评价方法将构造特定的评价指标。该指标的选择应该符合评价的主客观标准,采用定量的方式来评估结果模式集中有效的、新颖的、潜在可用的及最终可理解的模式。所采用的评价指标为分类正确率、查全率、查准率、综合分类率及平均准确率等指标。

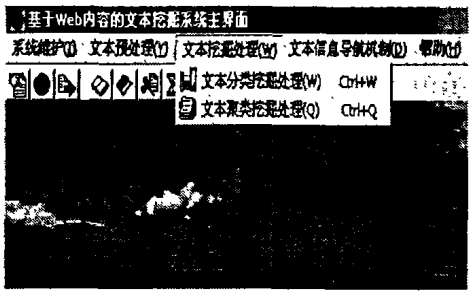

图6 Web 文本挖掘系统主界面(2)

(4)模式的解释与呈现:Web 文本挖掘系统最终挖掘出来的模式能够用可视化的方式进行显示,同时对用户提供概念导航机制的功能,使用户有效、快速地浏览和获取信息。

结论 通过分析,可以得出:

(1)UMKDSS 刻画了结构化数据挖掘与非结构化数据挖掘之间的关系,指出了复杂类型数据挖掘的重点和难点在于它的特征处理,即如何提取特征并降低特征空间的维数,使其近似转化为结构数据,然后再进行数据挖掘。

(2)UMKDSS 既包含结构化数据挖掘,又包含复杂类型数据挖掘,并将结构化数据挖掘与复杂类型数据挖掘有机地统一起来,成为知识发现领域的统一框架理论,对 KDD 的发展有积极的指导意义。

参考文献

1 李德毅. 发现状态空间理论[J]. 小型微型计算机系统. 15(11)
2 邸凯昌. 空间数据发掘与知识发现[M]. 武汉大学出版社,2001. 20～23
3 Yang Bingru. Double-Bases Cooperating Mechanism in KD(D&K) System[C]. IC-AI'99,1999 (USA)
4 Yang Bingru. Structrue of Knowledge Discovery Based on Double-Base Cooperating Mechanism[J]. Data Mining and Knowledge Discovery,1999
5 唐菁. 基于知识发现内在机理的 Web 文本挖掘结构模型与算法研究[D]. 北京:北京科技大学,2003