

# 一种基于粗糙集理论的 XML 数据挖掘模型

朱兴统 许 波

(广东石油化工学院计算机与电子信息学院, 茂名 525000)

**摘要** 随着大量的 XML 数据的出现, 给数据挖掘领域提出了新的挑战。传统数据挖掘是基于关系数据库和数据仓库的, 如何挖掘出 XML 形式的数据成为研究的热点问题。由于 XML 文档是一种半结构化数据, 使用传统的数据挖掘方法对 XML 数据进行挖掘是不适用的。提出了一种基于粗糙集理论的 XML 挖掘模型, 并进行了实验, 结果表明利用粗糙集理论对 XML 数据挖掘是可行的。

**关键词** XML 粗糙集 数据挖掘

**中图法分类号** TP391.3; **文献标示码** A

粗糙集理论是波兰数学家 Pawlak Z W 在 1982 年提出的一种分析数据的数学理论, 其主要思想是在保持分类能力不变的前提下, 通过知识约简, 导出问题的决策或分类规则<sup>[1,2]</sup>。

粗糙集理论的特点是不需要预先给定任何先验知识, 而是直接从给定问题的描述集合出发, 从中发现规则。经过二十多年的发展, 粗糙集理论和应用取得了很快的发展, 它被认为是一种具有极大潜力和有效的知识获取工具, 已成功应用于机器学习、决策分析、图像处理、医疗诊断、模式识别和数据挖掘等领域<sup>[3]</sup>。数据挖掘就是从大量的数据中, 提取人们感兴趣的知识, 这些知识是隐含的、事先未知的信息。提取的知识表示为概念、规则、规律和模式等形式。

近年来, 以 XML<sup>[4]</sup>为基础的新一代 Web 环境的出现, 很好地兼容了原有的 Web 应用, 而且可以更好地实现 Web 中的信息共享与交换。XML 在信息管理、电子商务、个性化出版、移动通信、网络教育、电子文档交换等诸多领域得到了广泛应用, 而且其应用范围还在不断扩展。对于这些越来越多的采用 XML 文档格式进行存储、交换和表现的数据, 除

了已有的信息抽取、Web 搜索等信息处理方法之外, 人们越来越需要获取更进一步的、深层次的知识, 这就需要对其进行数据挖掘。由于 XML 是一类半结构化数据, 而传统的数据挖掘技术主要面对的是以结构化数据为主的关系数据库、事务数据库和数据仓库, 不能直接将传统的基于关系数据库的挖掘方法用于 XML 数据的挖掘。因此, 研究出面向 XML 数据的挖掘方法成为数据挖掘领域的一项重要课题, XML 数据挖掘也是一个研究热点问题<sup>[5,6]</sup>。

## 1 XML 数据

XML(Extensible Markup Language)意为可扩展的标记语言, 用户可以定义自己的标记, 用来描述文档的结构。XML 是 W3C 在 1998 年制定的一项标准, 是标准通用标记语言的 SGML 的一个子集。XML 语言的规范性、灵活性和强大的语言描述能力, 能够满足对异构数据进行整合。XML 语言已经成为互联网上进行数据表示和数据交换的标准。XML 数据模型是一种类似于树结构的层次嵌套模型。构造 XML 文档的基本成分是元素(Element), 元素由标签(Tag)定义, 由起始标签、元素内容和结束标签组成。XML 文档的样式如下所示。

<? xml version = "1.0" encoding = "gb2312" ? >

2011 年 4 月 12 日收到

第一作者简介: 朱兴统(1974—), 男, 海南文昌人, 硕士, 讲师, 研究方向: 数据挖掘、计算智能。

```

< customers >
< customer id = "1" >
< sex > man </ sex >
< age > 40 </ age >
< income > low </ income >
< profession > blue </ profession >
< notebook > No </ notebook >
</ customer >
< customer id = "2" >
< sex > woman </ sex >
< age > 30 </ age >
< income > high </ income >
< profession > white </ profession >
< notebook > Yes </ notebook >
</ customer >
</ customers >

```

## 2 粗糙集理论相关基本概念

### 2.1 知识表达系统的定义

知识表达系统的基本成分是研究对象的集合,关于这些对象的知识是通过制定对象的基本特征(属性)和它们的特征值(属性值)来描述的,所以知识表达系统可以形式化表示为: $S = (U, A, V, f)$ ,其中:

$U$ :是一个非空有限对象(元组)集合,称为论域;

$A$ :为属性的非空有限集合;

$V = \cup a \in Va$ ,  $Va$  是属性  $a$  的值域;

$f: U \times A \rightarrow V$  是一个信息函数,它为每个对象的每个属性赋予一个信息值,即:  $Va \in A$ ,  $x \in U$ ,  $f(x, a) \in Va$ 。

知识表达系统也称为信息系统。通常也用  $S = (U, A)$  代替  $S = (U, A, V, f)$ 。

当  $A$  中的属性集可进一步分解为  $C$  和  $D$ ,且满足  $A = C \cap D$ ,  $C \cap D = \varphi$  时, $C$  称为条件属性集, $D$  称为决策属性集。具有条件属性和决策属性的知识表达系统称为决策表。

### 2.2 不可分辨关系

设  $U$  为一个有限的非空论域, $R$  为  $U$  上的一簇等价关系,若  $P \in R$ ,且  $P \neq \varphi$ ,则  $\cap P$ (所有的  $P$  中等

价关系的交集)也是一个等价关系,称为  $P$  上的不可区分关系,记为  $\text{ind}(P)$ ,且有

$$[x]_{\text{ind}(P)} = \cap [x]_R.$$

这样, $U/\text{ind}(P)$  表示与等价关系簇  $P$  相关的知识,称为知识库  $K = (U, R)$  中关于  $U$  的  $P$  基本知识( $P$  基本集)。

**定义 2** 上、下近似集 若  $X \subseteq U$ ,则称  $\underline{R}(X) = \{x \in U : [x]_R X\}$  为  $X$  的下近似集,  $\bar{R}(X) = \{x \in U : [x] \cap X \neq \varphi\}$  为  $X$  的上近似集。 $\bar{R}(X) = \{x \in U : [x] \cap X \neq \varphi\}$ 。

$\text{pos}_R(X) = \underline{R}(X)$  称为  $X$  的  $R$  正域, $\text{neg}_R(X) = U - \underline{R}(X)$  称为  $X$  的  $R$  负域。

### 2.3 知识约简与核

知识约简是粗糙集理论的核心内容之一。知识约简就是在保持知识库分类能力不变的条件下,通过消除不必要的知识,最终得到信息系统的分类或决策规则的方法。知识约简分为属性约简和属性值的约简。

**定义 3** 设  $R$  是一个等价关系族, $r \subseteq R$ ,如果  $\text{IND}(R) = \text{IND}(R - \{r\})$ ,则称  $r$  在  $R$  中是不必要的;否则称  $r$  在  $R$  中是必要的。

**定义 4** 如果任一  $r \subseteq R$  是  $R$  中必要的,则等价关系族  $R$  是独立的;否则  $R$  是依赖的。

**定义 5** 设  $Q \subseteq P$ ,若  $Q$  是独立的,并且  $\text{IND}(Q) = \text{IND}(P)$ ,则称  $Q$  是关系族集  $P$  的一个约简。在  $P$  中所有不可省的关系集合称为  $P$  的核,记为  $\text{CORE}(P)$ 。也就是说  $P$  的核等于  $P$  中所有约简的交集,即: $\text{CORE}(P) = \cap \text{RED}(P)$ ,其中  $\text{RED}(P)$  是  $P$  的所有约简的族集。

## 3 基于粗糙集理论的 XML 数据挖掘模型

基于粗糙集理论的知识获取,主要是通过将 XML 数据转换成决策表,然后对决策表进行约简,在保持决策表决策属性和条件属性之间的依赖关系不发生变化的前提下对决策表进行约简。决策表中的数据约简分为两部分,一是对属性进行约简,一是对属性值进行约简。

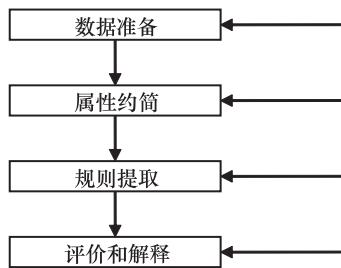


图 1 基于粗糙集理论的数据挖掘模型

### 3.1 数据挖掘过程

#### 3.1.1 数据准备

数据准备又可分为 3 个步骤:数据选取、数据预处理和数据变换。数据选取的目的是确定发现任务的操作对象,即目标 XML 数据是根据用户的需要从原始 XML 数据中抽取的一组数据。数据预处理主要是对目标 XML 数据进行再加工,检查数据的完整性以及数据的一致性,对其中的噪声数据进行处理。一般包括对数据缺省值的处理、数据的离散化等。数据变换的主要目的是削减数据维数或降维,即从初始特征中找出真正有用的特征以减少数据挖掘时要考虑的特征或变量个数。从经过上面处理过的 XML 数据中挑选出条件特征作为决策表的条件属性,把结果数据的特征作为决策属性。

#### 3.1.2 属性约简

利用区分矩阵从决策表中消去一些不必要的属性,使决策表得到简化。

#### 3.1.3 规则提取

规则提取的过程就是对决策表进行值约简的过程。属性值约简就是在属性约简的基础上,针对每一个决策规则,去除表达该规则的冗余属性值,以便进一步简化决策规则。经过属性值约简之后得到新的信息表,所有属性值均为该表的值核,所有记录均为该信息表的规则。

#### 3.1.4 评价与解释

根据最终用户的决策目的对提取的模式和规则进行分析,删除冗余或无关的规则,把最有价值的规则区分出来。如果模式不满足用户要求,就需要退回到前面的阶段。最后把提取的决策规则表

示成 XML 数据形式。

### 3.2 约简算法

区分矩阵是由波兰华沙大学的著名数学家 Skowron 提出来的,是近年来在粗糙集约简上出现的一个有力工具。利用这个工具,可以将存在于复杂的信息系统中的全部不可区分关系表达出来。

**定义 6** 区分矩阵由华沙大学数学家 Skowron 提出,有系统  $S = (U, A)$ , 其中  $A = C \cup D$ ,  $a(x)$  是  $x$  在属性  $a$  上的值, 区分矩阵  $M$  为:

$$(C_{ij}) = \begin{cases} a \in A : a(x_i) \neq a(x_j), D(x_i) \neq D(x_j); \\ 0, & D(x_i) = D(x_j). \end{cases}$$

**定义 7** 区分函数是从区分矩阵中构造的。约简算法是先求  $C_{ij}$  的每个属性的析取,然后再求所有  $C_{ij}$  的合取。

基于区分矩阵的属性约简主要思想是通过构造区分矩阵,并且化简由区分矩阵导出的区分函数,在使用吸收律化简区分函数成标准式后,所有的蕴含式包含的属性就是信息系统的所有约简集合,最后选取包含核属性的最小约简集作为最终约简集。用区分矩阵不但可以求得属性约简,同样也可以求属性值约简,这就大大简化了决策表规则提取的算法实现<sup>[7]</sup>。

属性约简算法描述:

设  $M$  是决策表  $T$  的可辨识矩阵,  $A = \{a_1, a_2, \dots, a_n\}$ , 是  $T$  中所有条件属性的集合。 $S$  是  $M$  中所有属性组合的集合,且  $S$  中不包含重复项。令  $S$  中包含有  $s$  个属性组合,每个属性组合表示为  $B_i$ ,其公式化描述为  $B_i \in S, B_j \in S, B_i \neq B_j (i, j = 1, 2, \dots, s)$ 。设  $\text{Card}(B_i) = m$ , 则  $B_i$  中每个条件属性表示为  $b_{i,k} \in B_i (k = 1, 2, \dots, m)$ 。Redu 是决策表  $T$  属性约简后得到的属性集合。

算法具体步骤如下。

第 1 步将核属性列入属性约简后得到的属性集合,即  $\text{Redu} = C_0$ ;

第 2 步在区分矩阵中找出所有不包含核属性的属性组合  $S$ , 即

$$Q = \{B_i, B_i \cap \text{Redu} \neq \emptyset, i = 1, 2, \dots, s\}, S = S - Q;$$

第 3 步将属性组合  $S$  表示为合取范式的形

式,即

$$P = \bigwedge \{ \forall b_{i,k} : (i=1,2,\dots,s; k=1,2,\dots,m) \};$$

第4步将  $P$  转化为析取范式形式;

第5步根据需要选择满意的属性组合。如需属性数最少,可直接选择合取式中属性数最少的组合。

## 4 实验结果与分析

本文使用 Java 实现了基于粗糙集理论的 XML 数据挖掘算法,使用 JDOM<sup>[8]</sup>访问 XML 数据,实验的 XML 数据由某电器超市的交易数据库转换成 XML 文档的格式,共有 1000 个会员客户的数据。数据中包含性别、年龄、职业、收入和是否购买笔记本电脑,具体属性及值域如下所示。

条件属性 a,b,c 和 d 及值域:

a:性别 值域:1:男 2:女;

b:年龄 值域:1:青年 2:中年 3:老年;

c:职业 值域:1:白领 2:蓝领 3:没工作;

d:收入 值域:1:( $\geq 5000$  元) 2:(2 000 ~ 4 999元) 3:( $< 2000$  元)。

决策属性 e,值域:1:购买笔记本电脑 2:未购买笔记本电脑。

通过对 1 000 条顾客数据进行挖掘,得到的规则也表示成 XML 文档格式,部分规则表示如下:

```
<? xml version = "1.0" encoding = "gb2312"? >
```

```
<rules>
<rule>
<antecedent>
<a>1</a>
<b>1</b>
<c>1</c>
<d>1</d>
</antecedent>
<consequent>
<e>1</e>
</consequent>
</rule>
<rule>
<antecedent>
```

```
<c>1</c>
<d>2</d>
</antecedent>
<consequent>
<e>1</e>
</consequent>
</rule>
<rule>
<antecedent>
<c>3</c>
<d>3</d>
</antecedent>
<consequent>
<e>2</e>
</consequent>
</rule>
.....
</rules>
```

通过与一个面向关系数据库的粗糙集挖掘算法结果对比,挖掘结果完全一致,证明基于粗糙集理论的 XML 数据挖掘算法是正确的。本文算法只是针对一个 XML 文档,且 XML 数据的结构是符合特定的结构,算法还不能对任意的 XML 文档进行挖掘。在处理大的 XML 文档时,JDOM 可能会受到内存问题的困扰。

## 5 结束语

由于 XML 文档是一种半结构化数据,使用传统的数据挖掘方法对 XML 数据进行挖掘是不适用的。本文提出了一种基于粗糙集理论的 XML 挖掘模型,经实验结果表明利用粗糙集理论对 XML 数据挖掘是可行的。XML 是正在发展的技术,对 XML 数据挖掘本身还有许多技术有待进一步完善,本算法为深入研究针对 XML 数据挖掘提供了借鉴。

## 参 考 文 献

- 1 Pawlak Z. Rough set, International Journal of Computer and Information Sciences, 1982;11(5):341—356
- 2 Pawlak Z. Rough set theory and its applications to data analysis. International Journal of Cybernetics and Systems, 1998; 29 (7): 661—688

- 3 王国胤,姚一豫,于 洪. 粗糙集理论与应用研究综述. 计算机学报,2009;32(7):1229—1240
- 4 W3C. Extensible Markup Language (XML) 1.0 (Fifth Edition),<http://www.w3.org/TR/2008/REC-xml-20081126/>
- 5 屈志毅,周海波,马晓军,等. 决策树在 XML 数据库挖掘中的研究. 计算机工程与设计,2008;29(14):3363—3368
- 6 杨 科,赖朝安,赵 阳. 基于 XML 数据的 FP-growth 算法挖掘研究. 计算机工程与应用,2008;44(19):150—159
- 7 常犁云,王国胤,吴 渝. 一种基于 Rough Set 理论的属性约简及规则提取方法. 软件学报,1999;10(11):1206—1211
- 8 哈罗德,刘文红. Java 语言与 XML 处理教程:SAX,DOM,JDOM,JAXP 与 TrAX 指南. 北京:电子工业出版社,2003

## A Model of Data Mining for XML Based on Rough Sets

ZHU Xing-tong, XU Bo

(School of Computer and Electronics Information, Guangdong University of Petrochemical Technology, Maoming 525000, P. R. China)

**[Abstract]** With the emergence of of XML data, the field of data mining raise new challenges. Traditional data mining is based on relational database and data warehouse, how to dig out the data as XML is a hot issue. The XML document is a semi-structured data, the traditional method of XML data mining data mining is not applicable. A rough set theory is proposed based on XML mining model, and experimental results show that rough set theory using XML data mining is feasible.

**[Key words]** XML    rough sets    data mining

(上接第 4894 页)

- 2 王功明,郭新宇,赵春江,等. 基于 Koch 曲线的土壤孔隙三维可视化仿真. 系统仿真学报,2008;20(3): 662—668
- 3 李水根. 分形. 北京:高等教育出版社, 2004
- 4 孙博文. 分形算法与程序设计. 北京:科学出版社, 2004

## New Design of Recursion Algorithm for a Class of Fractal Curves

TONG Ning-jiang<sup>1</sup>, GU Hui<sup>2\*</sup>

(Department of Mechanical Electrical Engineering, Taizhou Vocational College of Science and Technology<sup>1</sup>, Huangyan 318020, P. R. China;  
College of Information Engineering, Zhejiang University of Technology<sup>2</sup>, Hangzhou 310032, P. R. China)

**[Abstract]** a class of fractal curves is named Koch structure. A popular algorithm for generating Koch structure is recursion algorithm. About Koch structure, aims at the limitation of current recursion algorithms, ordinal number theory and some properties are brought up. Based on this, designs a new recursion algorithm is designed, two realizations of new algorithm are given, new solutions to generate the 2D Koch structure is offered. New algorithm can be extended to 3D space, effectively solves the problem of general constructing 3D Koch structure.

**[Key words]** fractal    Koch structure    recursion algorithm