

一种面向 XML 文档的模糊关联规则算法

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。传统数据挖掘算法是面向关系数据库和数据仓库的,不能直接用于 XML 文档的数据挖掘。从模糊集的基本理论入手,通过定义模糊概念软化属性论域的划分边界,提出了一种面向 XML 数据的模糊关联规则挖掘方法,并且使用 Java 语言实现,实验结果证明算法是正确的。

关键词 XML 文档 数据挖掘 模糊关联规则

中图法分类号 TP311.138; **文献标志码** A

随着 Internet 的普及和 Web 技术的快速发展,作为 Internet 上信息表示和数据交换标准的 XML 应运而生。XML 在电子商务、电子数据交换、科学数据表示、数据建模与分析 and 搜索引擎等领域有着广泛的应用,而且其应用范围还在不断扩展。XML 数据的广泛应用不仅给 Web 带来了更鲜活的生命力,也给计算机领域研究工作带来了新的挑战。XML 技术吸引了数据库、多媒体、网络安全、信息检索等诸多领域研究者的关注。到目前为止,在 Web 上已经积累了大量的 XML 文档数据,因此,研究出有效的针对 XML 文档的数据挖掘方法成为数据挖掘领域和 XML 技术领域的一项重要课题。

数据挖掘(Data Mining)是从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。关联规则是数据挖掘中最先研究的对象之一,也是数据挖掘的一个主要研究方向。关联规则是由 Agrawal、Imielinski 和 Swami 等人在 1993 年首先提出的^[1],起初是研究顾客交易数据库中购买商品之间的关联规则的挖掘问题。1994 年, Agrawal 和 Verkamo 提出了关联规则挖掘的经典算法 Apriori^[2]。在处理数量型数据挖掘时,为了解决

“边界划分过硬”的问题,模糊集能较好地实现连续量和离散量之间的转换,并最终软化了属性论域的划分边界。本文将研究实现面向 XML 文档的模糊关联规则算法。

1 相关知识

1.1 XML 数据

XML 即可扩展标记语言(Extensible Markup Language),是一种与平台无关的标识数据的方法。XML 同 HTML 一样,都来自 SGML(标准通用标记语言)。SGML 是一种在 Web 发明之前就早已存在的用标记来描述文档资料的通用语言。但由于 SGML 语言过于庞大和复杂,不便于学习和使用,便有了 HTML 的诞生。伴随着 Web 应用的深入发展,HTML 也不能满足广泛的需求,于是便又有了 XML 的产生。XML 与 SGML 一样,是一个用来定义其他语言的元语言。与 SGML 相比,XML 规范简单易懂,是一种既无标签集也无语法的新一代标记语言。

XML 文档的逻辑结构由以下几个部分组成。

- (1) XML 文档通常以一个 XML 声明开始。
- (2) 通过 XML 元素来组织 XML 数据。
- (3) 为了组织数据能方便、清晰,XML 在字符数据中引入 CDATA 数据块。
- (4) 在文档中引入注释。

2011 年 6 月 7 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。E-mail: zhu898@126.com。

(5) 需要给 XML 处理程序提供一些指示信息, XML 文档中可以包含处理指令。

一个简单的 XML 文档如下所示:

```
<? xml version = "1.0" encoding = "gb2312" ? >
<students >
  <student id = "1" >
    <programing >80 </programing >
    <database_system >90 </database_system >
    <java >82 </java >
    <data_structure >75 </data_structure >
    <operating_system >64 </operating_system >
  </student >
  <student id = "2" >
    <programing >82 </programing >
    <database_system >76 </database_system >
    <java >85 </java >
    <data_structure >65 </data_structure >
    <operating_system >72 </operating_system >
  </student >
</students >
```

1.2 模糊集合理论

1965 年, Zadeh 教授发表论文“模糊集合”(Fuzzy set), 标志模糊数学的诞生^[3]。模糊集合就是指具有某个模糊概念所描述的属性的对象的全集。由于概念本身不是清晰的、界限分明的, 因而对象对集合的隶属关系也不是明确的、非此即彼的。人们的思维中还有着许多模糊的概念, 例如年轻、很大、暖和、傍晚等, 这些概念所描述的对象属性不能简单地用“是”或“否”来回答。模糊集合的基本思想是把经典集合中的绝对隶属关系灵活化, 即元素对“集合”的隶属度不再是局限于取 0 或 1, 而是可以取从 0 到 1 间的任一数值。比如“老人”是个模糊概念, 70 岁的肯定属于老人, 它的从属程度是 1, 40 岁的人肯定不算老人, 它的从属程度为 0, 55 岁属于“老人”的程度为 0.5, 60 岁属于“老人”的程度 0.8。

模糊集合中的特征函数, 被称为“隶属函数”。用隶属函数来刻画处于中间过渡的事物对差异双方所具有的倾向性。给定论域 X 上的一个模糊集合 A , 对任意 $x \in X$, 都有确定的一个数 $\mu_A(x)$, 且 $0 \leq \mu_A(x) \leq 1$, 其中: $\mu_A(x)$ 表示 x 对 A 的隶属度,

$\mu_A(X)$ 称为 A 的隶属函数。隶属函数是模糊理论中的重要概念, 正确地确定隶属函数是恰当地表现模糊概念的基础。常用的模糊隶属函数有梯形或半梯形分布、抛物线型分布、正态分布、高斯分布、钟型函数等。梯形隶属函数如下所示, 由四个参数 $\{a, b, c, d\}$ 确定。

$$f(x, a, b, c, d) = \begin{cases} 0, & x \leq a; \\ \frac{x-a}{b-a}, & a \leq x \leq b; \\ 1, & b \leq x \leq c; \\ \frac{d-x}{d-c}, & c \leq x \leq d; \\ 0, & d \leq x. \end{cases}$$

2 模糊关联规则

模糊关联规则挖掘^[4-6]是将模糊集合理论用在关联规则挖掘中, 主要用来处理数值型属性的区间划分问题, 用边界模糊的区间代替边界确定的区间。在模糊关联规则挖掘中, 每个模糊概念是一个模糊项, 属性上定义的所有模糊概念构成了全体模糊项的集合。设 t 是数据库 D 的一条记录, $t[A]$ 表示 t 在属性 A 上的取值。如果 x 是属性 A 的一个模糊项, 那么 t 对 x 的隶属度为: $\mu_x(t) = \mu_x(t[A])$ 。

定义 1 由一个或者多个模糊项组成的集合称为模糊项集, 项的个数为 k 的模糊项集 X 称为 k 阶模糊项集, 记为 $\{X[1], X[2], \dots, X[k]\}$, 其中 $X[i]$ 是在不同属性上定义的模糊项。 t 对 X 的隶属度定义为

$$\mu_x(t) = \prod_{i=1}^k \mu_x[i](t)。$$

定义 2 模糊项集 X 的支持数定义为 D 中所有记录对 X 的隶属度之和, 即 $X.\text{sup} = \sum_{i=1}^n \mu_x(t)$, 其中 n 为 D 的记录个数。模糊项集 X 在 D 中的支持度就是对模糊集 X 支持数与 D 的总记录数之比, 即 $F\text{Sup}(X) = \frac{X.\text{sup}}{n}$ 。

定义 3 对于给定的最小支持度 minsup , 如果模糊集 X 的支持度 $F\text{Sup}(X) > \text{minsup}$, 那么称 X 是大模糊项集。

定义 4 模糊关联规则“ $X \Rightarrow Y$ ”的模糊支持度定义为

$$FSup = \frac{\sum_{i=1}^n (\prod_{k=1}^p \mu x[i](t) \times \prod_{j=1}^q \mu y[j](t))}{n}。$$

定义 5 模糊关联规则“ $X \Rightarrow Y$ ”的模糊置信度定义为 $Fconf = \frac{FSup}{FSup(X)}$ 。

模糊规则的挖掘算法的主要思想:首先通过计算支持度来找到最大频繁模糊项目集,再从最大频繁模糊项目衍生出规则候选模糊集,最后对规则候选模糊集计算置信度,从而得到最后的模糊关联规则。模糊关联规则挖掘算法描述如下^[5]:

输入:数据库 $D = \{t1, t2, \dots, tn\}$,最小支持度,最小置信度。

输出:模糊关联规则

- (1) 将数量型属性离散化,并将记录在数量型属性上的取值划分成若干个模糊集等级。
- (2) 通过数据库 $D = \{t1, t2, \dots, tn\}$ 构造一个新的数据库,新数据库以数量型属性不同的模糊集等级作为数据库的模糊属性。
- (3) 在新数据库中计算所有 1-模糊属性集的模糊支持度,得到所有的 1-模糊频繁属性集。
- (4) 组合 1-模糊频繁属性集,得到 2-模糊候选属性集。
- (5) 计算所有 2-模糊候选属性集的模糊支持度,删除小于最小支持度的 2-模糊候选属性集,得到所有 2-模糊频繁属性集。
- (6) 组合第一个模糊属性相同的 2-模糊频繁属性集,得到 3-模糊候选属性集。
- (7) 查看 3-模糊候选属性集的子集:2-模糊属性集,删除含有不是 2-模糊频繁属性集的 3-模糊候选属性集,计算剩余 3-模糊候选属性集的模糊支持度,删除小于最小支持度的 3-模糊候选属性集,得到所有的 3-模糊频繁属性集。
- (8) 以此类推,直到发现所有的 k -模糊频繁集。
- (9) 从所有的模糊频繁集中生成不小于用户给定的最小置信度的模糊关联规则。

3 面向 XML 文档的模糊关联规则的实现

3.1 模糊化^[6]

为学生课程成绩定义三个模糊区间 { High, Middle, Low },隶属函数采用梯形隶属函数,成绩模糊隶属函数图形如图 1 所示。

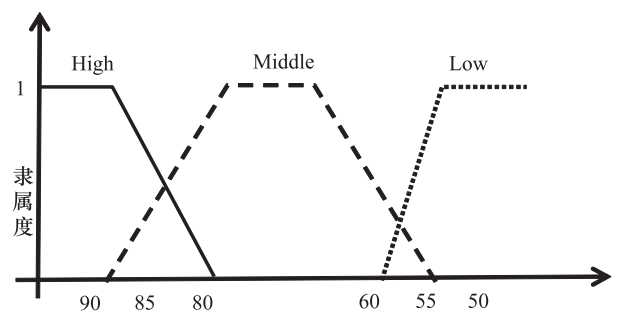


图 1 成绩隶属函数

经模糊化后的部分 XML 文档片段如下:

```
<? xml version = "1.0" encoding = "gb2312"? >
<students >
  <student id = "1" >
    <programing >
      <high >0 </high >
      <middle >1 </middle >
      <low >0 </low >
    </programing >
    <database_system >
      <high >1 </high >
      <middle >0 </middle >
      <low >0 </low >
    </database_system >
    <java >
      <high >0.2 </high >
      <middle >0.6 </middle >
      <low >0 </low >
    </java >
    <data_structure >
      <high >0 </high >
      <middle >0.2 </middle >
      <low >0.4 </low >
    </data_structure >
    <operating_system >
```

```

< high >0 </high >
< middle >0 </middle >
< low >0.6 </low >
</operating_system >
</student >
.....
</students >

```

3.2 算法实现与结果分析

本文使用 Java 语言开发,使用 JDOM^[7] 访问 XML 文档,实现了面向 XML 文档模糊关联规则算法,实验的 XML 文档数据是从我校学生成绩管理数据库中把 2002 ~ 2007 级计算机专业学生成绩转换成 XML 文档的格式,总共有 500 名学生的成绩记录,包含离散数学,高级语言程序设计,数据结构,操作系统原理与 Linux,数据库原理,计算机网络,面向对象原理与 Java 实践,计算机组成原理,单片机与接口技术,编译原理,软件工程等 11 门计算机专业主干课程的成绩。给定最小支持度为 0.03 和最小置信度为 0.75,得到的规则也表示成 XML 文档格式,部分关联规则表示如下:

```

< ? xml version = "1.0" encoding = "gb2312" ? >
< rules >
  < rule >
    < antecedent >
      < programing > high </programing >
    </antecedent >
    < consequent >
      < java > high </java >
    </consequent >
  </rule >
  < rule >
    < antecedent >
      < programing > high </programing >
      < data_structure > middle </data_structure >
    </antecedent >
    < consequent >
      < database_system > high </database_system >
    </consequent >
  </rule >
  .....
</rules >

```

本算法实现挖掘得到的结果与直接利用模糊关联规则对关系数据库中的学生成绩挖掘结果对

比,两者挖掘结果完全一致,证明使用 Java 语言实现的面向 XML 文档的模糊关联规则算法是正确的。在 CPU 为 Athlon64 X2 5000 +、内存为 2 G 的 PC 机上运行本算法,最小置信度为 0.75,最小支持度与执行时间的关系如图 2 所示。

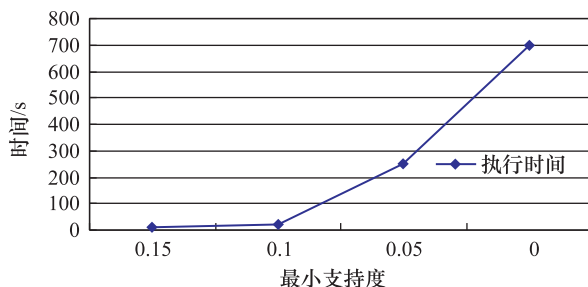


图2 不同的最小支持度与运行时间的关系

4 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。本文提出了利用 Java 语言实现模糊关联规则算法对 XML 文档数据进行挖掘,经实验结果表明是可行的。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。当数据属性很多时,模糊化生成的模糊 XML 文档也会很大,使用 JDOM 访问 XML 文档可能会受到内存问题的困扰。XML 是正在发展的技术,对 XML 文档数据挖掘本身还有很多问题有待进一步研究,本算法为深入研究针对 XML 数据挖掘提供了一点借鉴。

参 考 文 献

- 1 Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. Washington, USA;1993 ACM SIGMOD Conf, 1993
- 2 Agrawal R, Srikant R. Fast algorithms for mining association roles. Proceeding of the 20th International Conference on Very Large Databases,1994;(2):478—499
- 3 陈水利,李敬功,王向公.模糊集理论及其应用.北京:科学出版社,2005

(下转第 6477 页)

28—29

4 任 侠,吕述望. ARP 协议欺骗原理分析与抵御方法. 计算机工程,2003;9:127—128

5 黄睿达. 校园中 ARP 病毒原理和防范措施. 软件导刊,2010;(03):118—119

Design and Realization of the ARP Cheat Preventio in Campus Network

XIA Dong-liang, LIAO Meng-yi
(Pingdingshan University, Pingdingshan 467000, P. R. China)

[Abstract] the ARP cheat does harm to the normal function of the campus network. The principle and the common types of the ARP cheat are illustrated, and presented solutions. ARP cheat could be prevented by the port security of switches, functions such as ARP-Cheat and GSN® mechanism, which results in good practical effect.

[Key words] ARP cheat MAC address switch GSN mechanism



(上接第 6470 页)

4 吴君辉,殷肖川,张 薇. 基于模糊关联规则挖掘改进算法的 IDS 研究. 计算机测量与控制,2009;17(11):2256—2259

5 陆建江,徐宝文,邹晓峰. 模糊规则发现算法研究. 东南大学学报(自然科学版),2003;33(3):272—274

6 亢海力,王来生,蔡永旺. 基于概率的模糊加权关联规则挖掘. 计算机应用,2006;26(6):113—114

7 哈罗德,刘文红 编. 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 Fuzzy Association Rules Algorithm for XML Document

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 a large number of XML data, the field of data mining raises new challenges. Traditional data mining algorithm is oriented relational database and data warehouse, and can not be directly used for data mining in XML documents. From the basic theory of fuzzy sets, by defining the softening properties of the domain partition boundary, a fuzzy association rules oriented XML data mining is proposed, and implement it used the Java language. Experimental results show that the algorithm is correct.

[Key words] XML documents data mining fuzzy association rules