基于本体的关联规则在电子商务中的应用

吕 刚[1,2,3],郑 诚[1,2]

(1.安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039;
2.安徽大学 计算机科学与技术学院,安徽 合肥 230039;
3.合肥学院 计算机科学与技术系,安徽 合肥 230022)

摘 要:通过Web进行商务活动带来的便利和它所产生的交易速度已成为电子商务迅猛发展的关键推动力。对电子商务交易日志文件的挖掘可以得到关于群体客户行为和方式的普遍知识,从而改进服务,结合领域知识的关联规则挖掘问题一直是研究热点,通过利用领域本体整合关联关系提高挖掘效果。结合AROS算法,实验表明得到的规则更有意义。

关键词:本体;关联规则;AROS算法

中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)06-0250-03

Association Rules with Ontological Information in E-Commerce

LÜ Gang[1,2,3], ZHENG Cheng[1,2]

(1. Ministry of Education Key Laboratory of Intelligent Computing & Signal Processing, Anhui University, Hefei 230039, China;
2. School of Computer Science and Technology of Anhui University, Hefei 230039, China;
3. Department of Computer Science and Technology, Hefei University, Hefei 230022, China)

Abstract: The commercial activites carried out through the Web and bring it to facilitate the rapid development of e-commerce transaction speed has become the key driving force. Can get the knowledge of behavior and manner of the customer groups through the Web log mining. Mining association rules incorporated with domain knowledge has been studied recently, which can improve the mining results. The efficiency and advantages of this algorithm has been approved by experimental results.

Key words: ontology; association rule; AROS algorithm

0 引 言

目前,通过Web进行商务活动带来的便利和它所产生的交易速度已成为电子商务迅猛发展的关键推动力。对电子商务交易日志文件的挖掘可以得到关于群体客户行为和方式的普遍知识,从而改进服务,给客户个性化的界面,开展有针对性的电子商务以更好地满足访问者的需求,扩大商机。

关联规则挖掘(Association Rule Mining)是数据挖掘研究的一个重要分支,其目的是为了发现交易数据库中不同商品之间的联系规则,可以用来指导商家科学地安排进货、库存以及货架设计等,在电子商务领域可以用作商品推介等。关联规则是指包含了一组对象间待定关联关系的规则[1]。表达方式是$A \Rightarrow B$,经典的应用是购物篮分析问题。例如,有40%的顾客在购买"PC"的同时购买了"Flash Disk"。但是,在运行的数据库里并没有如:"PC"、"Flash Disk"等这样的抽象概念。取而代之的是这些产品的品牌名,比如:"Lenovo"、"Netac"。基于这样数据库的挖掘效果不好。主要原因是[2]:获得的大量规则没有实际意义,或者获得的规则太具体,很难掌握整体的情况[3]。

利用领域知识或背景知识,可在高层次上进行数据挖掘,产生高层次或多层次的规则,甚至是具有语义意义的规则,这些规则由高层次的抽象概念组成。

收稿日期:2008-09-17;**修回日期**:2008-12-04

基金项目:安徽省自然科学基金项目(050420204);安徽省高校自然科学研究项目(2006kj055B);合肥学院科研基金项目(08KY033ZR)

作者简介:吕 刚(1978-),男,讲师,硕士研究生,研究方向为数据挖掘;郑 诚,副教授,博士,研究方向为数据库、数据挖掘。

1 本体(Ontology)

1.1 基本概念

1993年,Gruber给出了Ontology的一个最为流行的定义,即"Ontology是概念模型的明确的规范说明"。Perez等人认为Ontology可以按分类法归纳出5个基

本的建模元语(Modeling Primitive)。这些元语分别为:类(classes),关系(relations),函数(functions),公理(axioms)和实例(instances)。

Ontology 的研究与应用主要包括 3 方面[1]:

(1) 理论上的研究,主要研究概念及其分类、Ontology 上的代数;

(2) 在信息系统中的应用,主要包括处理信息组织、信息检索和异构信息系统互操作问题;

(3) Ontology 作为一种能在知识层提供知识共享和重用的工具在语义 Web 中的应用。

数据挖掘是一个人机交互、不断重复的过程,专家的领域知识或背景知识的应用对挖掘过程具有补充和促进作用,经常用作引导发现过程以避免无意义的结果,利用本体可以很好解决这个问题[4,5]。

通过本体的信息之间的关系帮助选择合适的信息,可以减少假设空间,使输出结果更能理解,改善挖掘效果[6]。通过本体里语义之间的链接描述了事物之间的层次关系,从而加快数据挖掘过程,提高获取知识的效率和质量。

1.2 表示方式

本体可以用 RDF 资源描述框架(Resource Description Framework)表示。RDF 基于用 Web 标识符(称作统一资源标识符,Uniform Resource Identifiers 或 URIs)来标识事物,用简单的属性(property)及属性值来描述资源。这使得 RDF 可以将一个或多个关于资源的简单陈述表示为一个由结点和弧组成的图(graph),其中的结点和弧代表资源、属性或属性值。RDF 通过三元组表示资源:"主语"、"谓词"、"宾语",通过 XML 表示方式便于理解以及数据处理。这些表示方法对于电子商务的数据处理是非常重要的。例如"PC"和"Lenovo"的描述方法如下:

```
<rdfs:Class rdf:resource="PC"/>
<rdfs:Class rdf:about="Lenovo" rdfs:label="Lenovo">
<rdfs:subClassOf rdf:resource="PC"/>
</rdfs:Class>
<rdfs:Class rdf:about="Segate" rdfs:label="Segate">
<rdfs:subClassOf rdf:resource="PC"/>
</rdfs:Class>
```

这段代码表示的"PC"类是"Lenovo"父类。RDF 的查询语句可以 RDQL 语句,通过 HP 公司提供的插件 JENA 可以实现本体的父类和子类的查询。

1.3 本体构建

本体构建是一项十分复杂的系统工程,需要正确的开发思想指导和合适的开发工具辅助。可以从以下几个步骤进行:

(1)确定本体应用的目的和范围:这是建立本体的第一步,也就是所研究的领域或任务,建立相应的领域本体或过程本体。领域越大,所建本体就越大,因此需限制研究的范围。

(2)本体分析:定义本体所有术语的意义及其之间的关系,该步骤需领域专家的参与,对该领域越了解,所建本体就越完善。

(3)本体表示:一般用语义模型表示本体。

(4)本体检验:建立本体的基本标准是清晰性、一致性、完整性、可扩展性。清晰性就是本体中的术语应被无歧义地定义;一致性,也就是术语之间关系逻辑上应一致;完整性,本体中的概念及其关系也应是完整的,应包含该领域内所有概念,但往往很难达到,需不断完善;而可扩展性,即本体应该能够扩展,在该领域不断发展时能加入新的概念。

(5)本体的建立:对所建本体按以上标准进行检验,符合要求的可以文件形式存放,否则转(2)。

1.4 开发工具概述

文中采用 Java 来实现算法,虽然 Java 在运行效率方面稍微差一些,但是由于本算法大量的使用本体相关的知识,而相关的这些项目主要是使用 Java 编写的;另外 Java 具有比较好的可移植性,这样为进一步的应用打下了基础。具体来说文中主要使用了 HP 公司的开源项目 Jena 作为基础的本体处理引擎。

Jena 主要包括以下五个子模块:

(1)ARP:Jena 的 RDF/XML 编译器。

(2)持久化对象模块:可以以多种形式存储已经建立的模型。

(3)推理模块:用于支持 OWL 等规范。

(4)Ontology(本体)模块:用于支持 OWL,DAML+OIL 和 RDFS 等规范。

(5)RDQL 模块:一种用于 RDF 查询的语句。

2 领域本体实例

在图 1 中表示了采购项目"Sony VAIO"、"PC"、和"Desktop"都是泛化概念类,"Segate 60GB"和"RAM 512MB"都是子类。通过结合这个本体希望能够发现如同如下规则:

PC=>Sony VAIO 或 Segate 60GB=> Desktop PC

3 算　法

基于本体的多层次关联规则的挖掘分为三个子问

题：

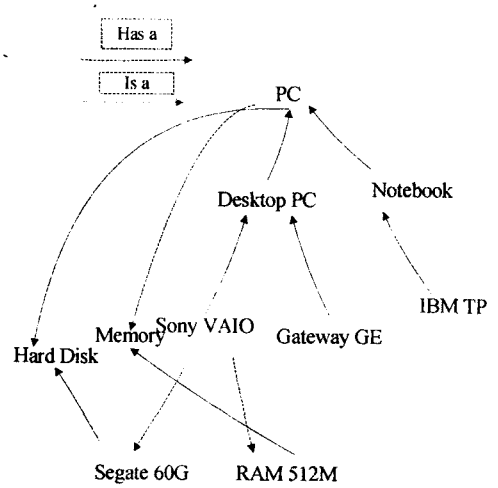

图 1 部分本体概念的语义关系

(1)找出事物数据库 D 中所具有用户指定最小支持度的项目集(itemset, I 的一个非空子集)。具有最小支持度的项目集称为频繁项目集(Fqerunet Itmeest)，反之就称为非频繁项目集。

(2) 利用频繁项目集生成所需要的多层次关联规则。对于频繁项目集 $ABCD$ 和 AB，如果比率 conf = support($ABCD$)/support(AB) 大于 minconf，就生成多层次关联规则 $AB \Rightarrow CD$ 。

(3)从所得到的多层次关联规则集中修剪非有趣规则。

问题(2)较为容易和直观，故关键集中在问题(1)和(3)上。

文中提出结合本体的 AROS (Association Rules with Ontological information Stratify algorithms) 算法[7,8]。AROS算法的特点是：统计发生的候选项目集，根据分类和合并关系的方法，将候选项目集 k - 选出项目集 C_k。MC_k 称为最大项目候选集，剩余项目集 RC_k。

详细的算法定义如下：

Output: L: the set of frequent itemsets for DB with T w.r.t. ms.

Steps:

1. repeat
2. if k=1 then Generate C_1 from item ontology T;
3. else C_k = apriori - gen(L_{k-1});
4. Delete any candidate in C_k that consists of classification or-composition relationship between items;
5. $MC_k = MC_{k-gen}(C_k, T)$;/* Using C_k, T to find maximal C_k */
6. Scan ED to count count(A) for each itemset A in MC_k;
7. $ML_k = \{A \mid A \in MC_k$ and sup(A) $\geqslant$ms$\}$;
8. $RC_k = RC_{k-gen}(C_k, MC_k, ML_k)$;/* Using C_k, MC_k, ML_k to find RC_k */
9. if $RC_k \neq$ then
10. Scan ED to count count(A) for each itemset A in RC_k;
11. $RL_k = \{A \mid A \in RC_k$ and sup(A) $\geqslant$ms$\}$;
12. end if
13. $L_k = ML_k \cup RL_k$;
14. until $L_k = \emptyset$
15. $L = \cup_k L_k$;

4 试验结果分析

提出基于本体的关联规则，解决在电子商务中的规则获取。通过 IBM 数据生成器产生的数据，进行试验认证。生成的数据有 362 个项目，每个交易有 20 个项目，30 类，4 层。设定 ms = 1.5%，试验结果表明结合本体的数据挖掘方法比普通的挖掘方法得到的规则更有意义，时间更短。普通的关联规则产生的结果数量十分巨大，这些结论完全由人来处理十分困难。基于本体的关联规则挖掘很好的解决了这个方面的问题，挖掘出来的结果更有概括性。

参考文献：

[1] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范 明，孟小峰译. 北京：机械工业出版社，2007.

[2] 袁万莲，郑 诚，翟明清. 一种改进的 Apriori 算法[J]. 计算机技术与发展，2008，18(5)：51 - 53.

[3] 陈 英，顾国昌. 基于领域本体的数据挖掘服务发现算法[J]. 计算机工程与应用，2008，44(18)：150 - 156.

[4] 邓志鸿，唐世渭，张 铭，等. Ontology 研究综述[J]. 北京大学学报：自然科学版，2002，38(5)：730 - 737.

[5] 邢平平，施鹏飞，赵 奕. 基于本体论的数据挖掘方法[J]. 计算机工程，2001，27(5)：15 - 16.

[6] Kuo Yen - Ting, Lonie A, Sonenberg L. Domain Ontology Drivern Data Mining[C]//Conference on Knowledge Discovery in Data archive Proceedings of the 2007 international workshop on Domain driven data mining table of contents. San Jose, California: [s.n.], 2007: 11 - 17.

[7] Kaya M, Alhajj R. Genetic algorithm based framework for mining fuzzy association rules[J]. Fuzzy Sets and Systems, 2005.152(3): 587 - 601.

[8] Hong T P, Kuo C S, Chi S C. Mining association rules from quantitative data[J]. Intell. Data Ana, 1999(3): 363 - 376.