地理信息关联规则挖掘算法的设计与应用*

吴卫华[1],袁宁[2]

(1.山东省智奥电算开发中心,山东 济南 250013;2.济南大学信息学院,山东 济南 250022)

摘要:本文着重就地理信息数据挖掘中的两种模式:关联规则和序列模式的概念和作用进行了探讨,阐述了在关联规则中寻找大项集算法的实现,以及在数据挖掘的序列模式基础上对寻找大项集算法的结果进行了改进和优化,使数据的关联规则与时间和序列之间建立了密切的联系,从而更好的实现了对于大规模地理信息数据库中数据的挖掘和利用。

关键词:关联规则;序列模式;地理信息

中图分类号:TP311.13 文献标识码:A

0 引言

地理信息系统是20世纪60年代以后迅速发展起来的一个新兴技术领域。它集计算机科学、地理学、测绘遥感学、空间科学、信息科学、环境科学和管理科学于一体,把图形管理系统和数据管理系统有机地结合起来,并且克服了图形系统和数据库系统各自固有的局限性,使二者优势更加突出。它能够对空间信息进行采集、存储、分析和表达,并能够处理其空间定位特征,把空间和属性信息有机地结合起来。基于自身强大的空间分析功能,地理信息系统已经广泛的应用在资源开发、环境保护、土地利用与规划、生态环境、市政管理、道路建设、石油地质等领域,越来越多的行业部门开始计划引进或自行开发地理信息系统。

地理信息系统中的数据是庞大和繁杂的,如何最有效的从中找到有用的信息和隐藏在数据内部的相互关联的规律,就需要利用数据挖掘的相关技术来实现。数据挖掘又称为数据库中知识发现,是一个从大量数据中抽取挖掘出未知的、有价值的模式或规律等知识的复杂过程。数据挖掘的主要任务就是从数据中发现模式[1]。其中关联模式是表示数据项之间的关联规则。关联规则挖掘就是从大量的数据中挖掘出有价值描述数据项之间相互联系的有关知识。序列模式与关联模式相仿,但是序列模式把数据之间的关联性与时间联系起来。为了发现序列模式,不仅需要知道事件是否发生,而且需要确定事件发生的时间。

1 关联规则

1.1 关联规则的概念及定义

挖掘数据的关联规则就是从给定的数据集中搜索数据项之间所存在的有价值联系。关联规则的形式如下:

设 $I=\{i_1,i_2,\cdots,im\}$ 为数据项集合,D 为与任务相关的数据集合。D 中的每个事务 T 是一个数据项子集。设 A 为一个数据项集合当且仅当 $A\subseteq T$ 时,称事务 T 包含 A。一个关联规则就是具有如下形式的一种蕴含式:$X\rightarrow Y$,其中 $X\subseteq I$,$Y\subseteq I$,且 $X\cap Y=\varnothing$。

()称数据集 X 具有大小为 s 的支持度,如果 D 中有 $s\%$ 的事务支持数据集 X;

()称关联规则 $X\rightarrow Y$ 在事务数据库 D 中具有大小为 s 的支持度,如果数据集 $X\cup Y$ 的支持度为 s;

()称规则 $X\rightarrow Y$ 在事务数据库 D 中具有大小为 c 的可信度,如果 D 中支持数据集 X 的事务中有 $c\%$ 的事务同时也支持物品集 Y。

事实上,人们一般只对满足一定的支持度和可信度的关联规则感兴趣。为了发现出有意义的关联

*收稿日期:2003-06-04;修订日期:2003-08-18;编辑:孟舞平

作者简介:吴卫华(969-),男,工程师,山东烟台人,主要从事地理信息系统工作。

规则,需要给定两个阈值:最小支持度和最小可信度[1]。前者即用户规定的关联规则必须满足的最小支持度,它表示了一组数据集在统计意义上的需满足的最低程度;后者即用户规定的关联规则必须满足的最小可信度,它反应了关联规则的最低可靠度。

1.2 关联规则的寻找大项集算法

利用大项集(itemsets)产生所需的规则(rules)。算法的思想在于:如果说 $ABCD$ 和 AB 是大项集,就可以通过计算可信度,也就是 conf = support($ABCD$)/support(AB),并通过 conf ≥ miniconf 来确定规则 $AB \rightarrow CD$ 是否确立(该规则由于 $ABCD$ 是大项集故肯定具有最小支持度)。

关联规则算法:

L_1 = {large 1-itemsets};//发现 1-项集

for ($k = 2$; $L_{k-1} \neq \phi$; $k++$) do begin

C_k = apriori-gen(L_{k-1});//根据频繁($k-1$)-项集产生侯选 k-项集

forall transactions $t \in D$ do begin

C_t = subset(C_k, t);//获得 t 所包含的侯选项集

forall candidates $c \in C_t$ do

c.count++;

end

L_k = {$c \in C_k$ | c.count ≥ minsup}

end Answer = $\bigcup_k L_k$;

其中 apriori-gen 函数以 L_{k-1}(所有大($k-1$)-项集)作为输入参数,返回所有大 k-项集的集合 L_k,由以下两步实现:

第一步,联合

insert into C_k

select p.item1, p.item2, …, p.item$_{k-1}$, q.item$_{k-1}$

from $L_{k-1}p$, $L_{k-1}q$

where p.item1 = q.item1, …, p.item$_{k-2}$ = q.item$_{k-2}$, p.item$_{k-1}$ < q.item$_{k-1}$;

第二步,剪枝(pruning),如果存在 c 的(k-1)-子序列不包含于 L_{k-1}之中,则删除所有项集 $c \in C_k$。

forall itemsets $c \in C_k$ do

forall ($k-1$)-subsets s of c do

if ($s \notin L_{k-1}$) then

delete c from C_k

2 序列模式

2.1 序列模式的概念及定义

序列模式就是发现随时间变化的交易序列(模式)。其分析的目的就是挖掘以发生时间顺序为主的项集发生序列[1]。如果一个交易包含一个项集,那么就称相应的一个交易序列包含一个项集序列,并满足以下条件:若交易序列中交易 j 包含项集序列中的第 I 个项集,那么交易序列包含项集序列中的第($+1$)个项集交易大于 j。项集序列的支持度就是包含它的交易序列(在整个交易序列中)所占比例。

2.2 一些相关概念及定义

数据源是一个给定的由客户交易(customer transaction)组成的大型数据库,每个交易(ransaction)由客户号(ustomer – id),交易时间(ransaction – time)以及在交易中购买的项(tem)组成。

()项集(temset)是由项(tem)组成的一个非空集合;

()序列(equence)是一列排好序的项集。

假定项集中的项由一些连续整数代替,这样一个项集 i 可以表示为($_1, i_2 \cdots i_m$),而这里的 i_j 代表了一个项[1]。一个序列 s 可以表示为 $< s_1, s_2 \cdots s_n >$,这里的 s_j 代表的是一个项集。

两个序列 $a < a_1, a_2 \cdots a_n >$ 和 $b < b_1, b_2 \cdots b_m >$,如果存在整数 $i_1 < i_2 < \cdots < i_n$ 且 a_1 包含于 b_{i1},a_2 包含于 b_{i2},…,a_n 包含于 b_{in},则称序列 a 包含于序列 b。比如序列 < ()()4,5)()> 包含于序列 < ()(,8)()4,5,6)()>,因为()包含于(,8),4,5)包含于 4,5,6)以及()包含于()。但是序列 < ()()> 不包含于 < (,5) >,反之亦然。前者表示项 3 和项 5 是先后购买的,而后者则表示项 3 和项 5 是同时购买的,这就是区别所在。在一个序列集中如果序列 s 不包含于任何其他序列中,则称序列 s 为最大的(maximal)。

一个客户所有的事务可以综合的看成是一个序列,每一个事务都由相应的一个项集来表示。事务按交易时间序排列就成了一个序列[1]。称这样的序列为客户序列(ustomer-sequence)。通常,将一个客户的交易按交易时间排序成 $T_1, T_2, \cdots\cdots, T_n$。$T_i$ 中的项集定义成 itemset(T_i)。这样,这个客户的客户序列就成了这样的一个序列:(temset(T_1) itemset

$(T_2)\cdots$itemset (T_n))。见表1。

表1 客户序列视图的数据库

Customer Id	Customer Sequence
1	< (30) (90) >
2	< (10,20) (30) (40,60,70) >
3	< (30,50,70) >
4	< (30) (40,70) (90) >
5	< (90) >

如果一个序列 s 包含于一个客户序列中,则称该客户支持序列 s。一个具体序列的支持定义为那一部分支持该序列的客户总数。

给定一个由客户交易组成的数据库 D,挖掘序列模式的问题就是在那些具有客户指定最小支持度的序列中找出最大序列。而每个这样的最大序列就代表了一个序列模式 (equential pattern)。

2.3 序列模式挖掘的步骤

下面根据序列模式挖掘原理,对由关联规则算法获得的大项集进行进一步的操作,分5个具体阶段来找出所有的序列模式。

2.3.1 排序阶段 (ort Phase)

数据库 D 以客户号 (ustomer-id)为主键 (ajor key),交易时间 (ransaction-time)为次键 (inor key)进行排序。实际上这个阶段将原来的事务数据库 (ransaction database)转换成由客户序列组成的数据库。

2.3.2 大项集阶段 (itemset Phase)

在这个阶段找出所有大项集组成的集合 L。也同步得到所有大1-序列组成的集合。因为这个集合就是 $\{<l> \mid l \in L\}$。

表2 大项集的映射

Large Itemsets	Mapped To
(30)	1
(40)	2
(70)	3
(40,70)	4
(90)	5

大项集被映射成连续的整数。在表1给出的数据库中,大项集分别是 (30), (40), (70), (40,70)和 (90)。表2给出了一个可行的映射。

这样映射的好处在于,将大项集按一个实体 (ntity)的形式进行处理,可以带来比较和处理上的方便和高效,提供了一个统一的格式。

2.3.3 转换阶段 (ransformation Phase)

在找序列模式的过程中,要不断地进行检测一个给定的大序列集合是否包含于一个客户序列中。为了使这个过程尽量的快,用另一种形式来替换每一个客户序列。

在转换完成的客户序列中,每条交易被其所包含的所有大项集所取代。如果一条交易不包含任何大项集,在转换完成的序列中它将不被保留。而如果一个客户序列不包含任何的大项集,在转换好的数据库中这个序列也将不复存在。但是在计算客户总数的时候,它仍将被计算在内。现在一个客户序列被一列由大项集组成的集合所取代,每个大项集的集合表示为 $\{l_1, l_2, \ldots, l_n\}$, l_i 表示一个大项集。

这样的一个转换好的数据库被称为 D_T。表2的数据库经过转换后在表3中得到了展示。比如,在对ID号为2的客户序列进行转换的时候,交易 (10,20)被剔除了,因为它并没有包含任何大项集;交易 (40,60,70)则被大项集的集合 {(40), (70) (40,70)}代替。

2.3.4 序列阶段 (equence Phase)

利用已知的大项集的集合来找到所需的序列。序列阶段算法的基本结构是对数据进行多次遍历。在每次遍历中,从一个由大序列组成的种子集开始,利用这个种子集,可以产生新的潜在的大序列。在遍历数据的过程中,计算出这些候选序列的支持度,这样在一次遍历的最后,就可以决定哪些候选序列是真正的大序列,这些序列构成下一次遍历的种子集。在第一次遍历前,所有在大项集阶段得到的具有最小支持度的大1-序列组成了种子集。

2.3.5 选最大阶段 (aximal Phase)

在大序列集中找出最大序列。在序列阶段找到所有的大序列之后,下述算法可以用来找出最大序列。定义最长序列的长度为 n,则:

for ($k = n$; $k > 1$; $k--$) do
foreach k-sequence sk do
Delete from S all subsequences of sk

表 3 转换后的数据库

Customer Id	Original Customer Sequence	Transformed Customer Sequence	After Mapping
1	< (30) (90) >	< {(30)} {(90)} >	< {1} {5} >
2	< (10,20) (30) (40,60,70) >	< {(30)} {(40), (70), (40,70)} >	< {1} {2,3,4} >
3	< (30,50,70) >	< {(30), (70)} >	< {1,3} >
4	< (30) (40,70) (90) >	< {(30)} {(40), (70), (40,70)} {(90)} >	< {1} {2,3,4} {5} >
5	< (90) >	< {(90)} >	< {5} >

3 关联规则挖掘在地理信息中的应用

在关联规则中寻找大项集算法的实现，以及在数据挖掘的序列模式基础上对寻找大项集算法的结果进行了改进和优化，使数据的关联规则与时间和序列之间建立了密切的联系，从而可以更好的实现对于大规模地理信息数据库中数据的挖掘和利用。

在建立土地管理信息系统中，利用数据挖掘的技术，实现了土地业务管理的科学化、规范化和信息化，提高了业务管理和决策的水平、质量和效率。以关联规则挖掘算法管理空间数据在数据维护的一致性、安全性、数据共享等方面都满足了土地管理的需要，同时结合关系数据库的海量数据管理、事务处理、并发控制等功能使用标准化的 SQL 语言对空间与非空间数据进行操作，使空间数据与非空间数据一体化集成，成为地理信息数据挖掘新的发展趋势。

4 结论

本文给出了对于数据挖掘中关联规则和序列模式的初步讨论，以及一些基本的概念和方法，并对于找大项集的问题进行了理论上探讨，对关联规则挖掘的算法所获得的结果在序列模式挖掘的基础上进行了进一步的操作，使数据的关联规则与时间和序列之间建立了密切的联系，从而更好的实现了对大规模事务数据库中有价值数据的挖掘工作。

参考文献：

[1] Rakesh Agrawal, Ramakrishnan Srikant . Mining Sequential Patterns. IBM Almaden Research Center, 1999, 73 – 84.

[2] Rakesh Agrawal, Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. IBM Almaden Research Center, 2001, 67 – 82.

[3] Sougata Mukherjea. WTMS: A Efficient Algorithm to Rank Web Resource. In Proc. of the 9th World Wide Web Conference, 2000, 55 – 97.

[4] Simon Fraser. Data Mining: Concept and Techniques Kluwer Academic Publishers, 2000, 167 – 187.

[5] 朱明. 数据挖掘 [M]. 北京: 中国科学技术大学出版社, 2002, 182 – 183.

Design and Application of Association Rules and Sequential Patterns in Geography Information Data Digging

WU Wei – hua[1], YUAN Ning[2]

(1. *Shandong GEO developing center*, *Shandong Jinan* 250013, *China*; 2. *School of information Science and Engineering*, *Jinan university*, *Shandong Jinan* 250022, *China*)

Abstract: As two models of geographic information data digging, association rules and sequential patterns are studied in this paper. Realization of Apriori algorithm in searching large item sets by using association rules is introduced as well. Based on sequential patterns of data digging, the result of apriori algorithm is optimized. Thus, a closed connection between association rules and time sequence is set up, whic will realize good use of large scale geographic information data base.

Key words: Association Rules; Sequential Patterns