数据源特性对序列模式挖掘结果的影响①

王翠青 陈未如 张 静 (沈阳化工学院 计算机科学与技术学院 辽宁 沈阳 110142)

摘 要： 用 IBM 数据生成工具生成数据考察了客户序列数、平均序列长度、平均事务长度、项目数等数据特性对挖掘结果序列模式个数的影响，得到了单项特性对挖掘结果序列模式个数的影响模型，用以在进行正式挖掘之前，通过数据特性对挖掘结果进行初步判断。

关键词： 数据挖掘；序列模式挖掘；数据特性

Influence of Data Source Features on Sequential Pattern Mining

WANG Cui-Qing, CHEN Wei-Ru, ZHANG Jing

(Shenyang University of Chemical Technology, Shenyang 110142, China)

Abstract: The article analyzes the influence of data source features on the number of sequetial parttern, which includes the number of customers, average transctions of sequences, average items of transctions, and number of different items. The model of single features influence is obtained. With the model, a primary jugdement can be got before mining.

Keywords: data mining; sequential pattern mining; data source features

序列模式挖掘是以支持度为基准进行挖掘。序列模式及在一定约束下的挖掘算法一直是学术讨论的重点，文献[1-3]提出了在不同领域应用的序列模式挖掘算法。而在序列模式挖掘中，支持度的设置直接影响着挖掘结果，在一次挖掘过程中，若支持度设置不合理，得到的挖掘结果就无法体现隐藏在数据源中的有用的信息。隐藏在数据源中的有用信息是由数据源的特性决定的，文章通过对数据源中客户序列个数、平均序列长度、平均事务长度、项目总数等数据特性在不同支持度下对挖掘结果的影响分析，找出各个数据特性对挖掘结果模式数目的影响程度，以此来指导实际挖掘中支持度的设定，及对挖掘结果进行预测。

1 实验基础

由于实际数据的数据特性很难进行把握，为了研究数据特性与挖掘结果的关系，排除数据特性间的相互干扰，本文实验数据采用 IBM QUEST 数据生成器进行生成。序列模式挖掘算法采用 PrefixSpan。

1.1 IBM Quest 数据生成器

IBM Quest 是做关联规则及序列模式挖掘多用的一种人工数据合成工具，这方面论文的实验数据大多是用它生成的数据。在其数据源生成程序中，共定义了 9 个属性。程序的主要特点是 9 个属性的所有数据均为均匀分布，并在程序的设计过程中加入了少量的判别条件以利于数据的合理性[4]。

对于生成的数据集 CmSnTkIi，约定 Cm 表示数据集的客户序列共有 m*1000 条；Sn 表示数据集的平均序列长度是 n；Tk 表示数据集中事务的平均长度是 k；Ii 表示数据集共有 i*1000 个项目。

1.2 PrefixSpan 算法

PrefixSpan 算法是 Han Jiawei 等人针对序列模式的类 Apriori 和 GSP 算法的缺点提出的基于投影的序列模式挖掘算法，该算法首先将原序列数据库转换成一系列的投影序列数据库，然后在这些投影序列数据库上挖掘出频繁序列模式[5]。该算法开放源代码，并且对序列模式挖掘的效率也较高。

① 基金项目:辽宁省教育厅科学研究计划(05L338)
收稿时间:2009-08-18;收到修改稿时间:2009-09-11

2　各数据特性对挖掘结果的影响

2.1 客户序列数

在仅有客户序列数变化的数据集上，以不同支持度进行挖掘，得到的序列模式个数如表 1-表 2 所示。

表 1　不同支持度下仅客户序列数变化的挖掘结果

	3%	1%	0.90%	0.70%	0.50%
C100S10T2.5I10	7	497	619	1000	1907
C80S10T2.5I10	7	494	622	1008	1899
C60S10T2.5I10	7	494	612	1006	1893
C40S10T2.5I10	7	491	616	1011	1919
C20S10T2.5I10	7	500	617	1006	1918
C5S10T2.5I10	7	489	635	1036	2047

表 2　不同支持度下仅客户序列数变化的挖掘结果(续)

	0.30%	0.10%	0.09%	0.07%
C100S10T2.5I10	5438	59928	72799	114032
C80S10T2.5I10	5414	60110	73166	110410
C60S10T2.5I10	5377	59090	71822	113621
C40S10T2.5I10	5540	61209	71553	110188
C20S10T2.5I10	5675	62024	75664	115980
C5S10T2.5I10	6695	93829	93829	171609

从表 1、表 2 可以看出，客户序列数对挖掘结果几乎没有影响，也就是说，在具备一定规模情况下，数据集的大小对挖掘不会产生决定性影响。因此，对于客户序列数十分庞大的数据源，可以通过对其子集进行预挖掘得到总体挖掘结果的概况。

2.2 平均序列长度

在数据特性 C10T2.5I10 固定，仅有平均序列长度变化的数据集上采用不同的支持度进行挖掘，在支持度一定的情况下，平均序列长度和序列模式数关系曲线如图 1 所示。

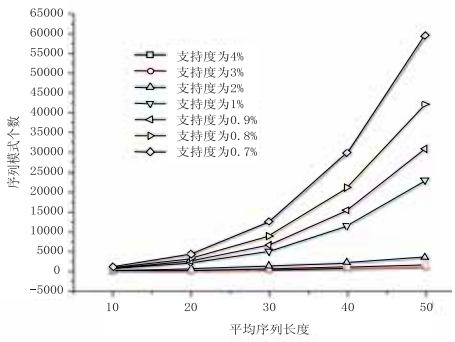

图 1　其他特性一定时不同支持度下平均序列长度和序列模式个数关系

考察在序列个数、平均事务长度、项目数、支持度相同的情况下，平均序列长度与序列模式个数关系，利用曲线拟合，得到两者关系可近似表示为 count=a×sb。图 2-图 3 为支持度为 0.7%和 2%时拟合结果。

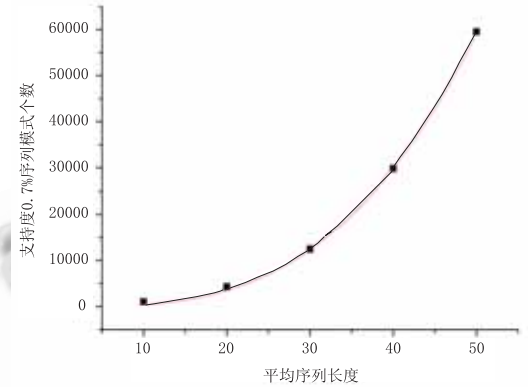

图 2　支持度为 0.7%时平均序列长度与序列模式个数关系曲线拟合结果

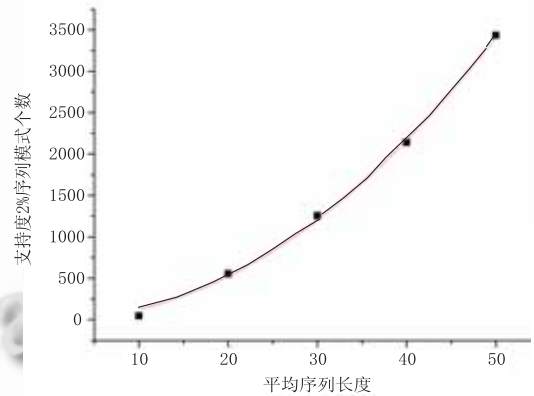

图 3　支持度为 2%时平均序列长度与序列模式个数关系曲线拟合结果

对 C10T2.5I10 一定，S 从 10 到 50 变化的所有数据集进行拟合，不同支持度下，参数 a 的值集中在 0.0709 到 0.4895 之间，b 的值集中在 2.0302 到 3.2319 之间，支持度越小 a 和 b 的值越大。实验表明在平均事务长度和平均项目数为其他值的情况下，此模型及变化趋势同样存在。

2.3 平均事务长度

在其他特性不变的情况下(C10S10I10)，平均事务长度与挖掘结果变化曲线如图 4 所示。

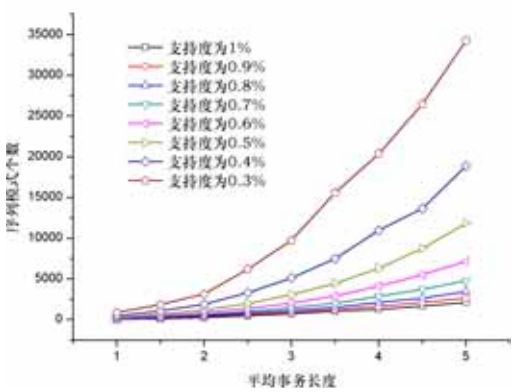

图 4 其他特性一定时不同支持度下平均事物长度与序列模式个数关系曲线

支持度一定时，平均事务长度越大，挖掘得到的序列模式个数越多。两者关系可通过 count=a×s2 得到相应拟合结果。支持度为 1%和 0.2%的拟合结果如图 5，图 6 所示。

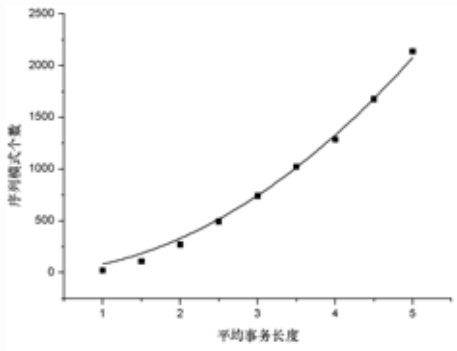

图 5 支持度 1%时平均事务长度和序列模式个数关系曲线拟合结果

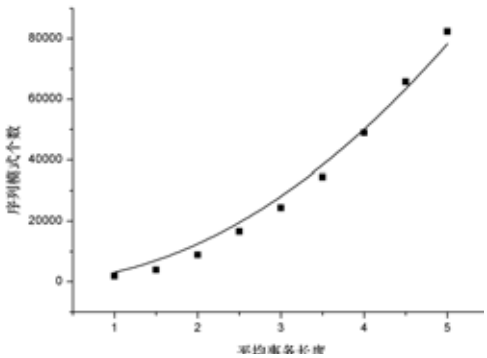

图 6 支持度 0.2%时平均事务长度和序列模式个数关系曲线拟合结果

对 C10S10I10 一定，T 从 1 到 5 变化的所有数据集进行拟合，不同支持度下，参数 a 的范围为 53.2 到 3133.1，支持度越小，参数 a 值越小。实验表明在平均事务长度和平均项目数为其他值的情况下，此模型及变化趋势同样存在。

2.4 项目数

在客户序列数、平均序列长度、平均事务长度不变，C10S10T25，项目数与支持度变化曲线如图 7 所示。

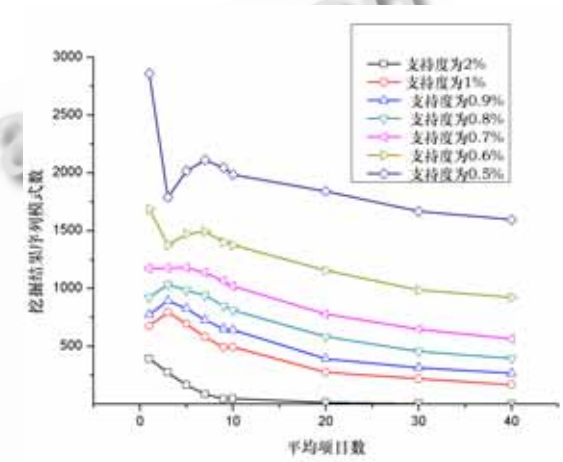

图 7 C10S10T10 不同支持度下平均项目数与挖掘结果关系

总体趋势体现为项目数越多，挖掘得到的序列模式个数越少。然而在支持度较小时，变化趋势会在平均项目数为 10 附近有所变化。另考虑一组数据(C10S7T4)，递减及波动趋势同样存在，如图 8 所示。

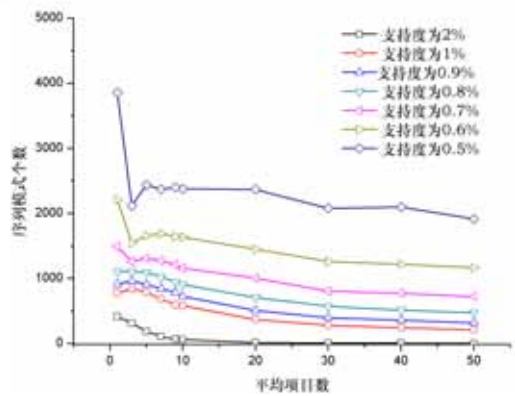

图 8 C10S7T4 不同支持度下平均项目数与挖掘结果关系

(下转第 18 页)

(上接第 193 页)

3 结论及进一步工作

在序列模式挖掘中，挖掘的结果必定由数据源特性决定，文章讨论了客户序列数、平均序列长度、平均事务长度及项目数等数据特性对挖掘结果的影响，并且找出了单项特性的影响程度。进一步工作考察平均项目数对结果影响的模型及此四种特性对挖掘结果的综合影响，更进一步的在挖掘前得到挖掘结果的预测值。

参考文献

1 冯林,于孝航,孙焘.基于最长公共子序列距离的主旨模式挖掘算法.计算机工程, 2008,34(14):47－48.

2 王淼,尚学群,薛贺.基于相邻模式段组合的生物序列模式挖掘算法.计算机工程与应用, 2008,2:194－197.

3 肖仁财,薛安荣.一种挖掘多维序列模式的有效方法.计算机工程与应用, 2008,6:191－194.

4 纪元,陈未如,张雪.并发关系模式合成数据源生成方法.山东大学学报(理学版), 2007,42:84－87.

5 Pei J, Han JW, Mortazavi AB, Pinto H. PrefixSpan: Mining sequential patterns efficiently by prefix2projected patterngrowth. In: Proc. of the 17th International Conference onData Engineering, Heidelberg, Germany, 2001. 215－226.