基于兴趣度的关联规则挖掘

李伟东,倪志伟,刘 晓
(合肥工业大学 管理学院,安徽 合肥 230009)

摘 要:关联规则挖掘是数据挖掘领域中的重要研究内容之一。然而,传统的基于支持度-可信度框架的挖掘方法可能会产生大量不相关、甚至是误导的关联规则。针对现有关联规则挖掘的评价标准存在的问题,提出在评价标准中增加兴趣度,并给出了兴趣度的定义和基于兴趣度的关联规则挖掘算法。利用兴趣度将关联规则分为正关联规则和负关联规则,从而可以用算法挖掘带有负项的关联规则。实验结果分析表明,在传统挖掘方法的基础上引入兴趣度,可以有效地减少正关联规则的规模,产生有意义的负关联规则。

关键词:关联规则;负关联规则;兴趣度
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2007)06-0080-03

Mining Association Rules Based on Interest Measure

LI Wei-dong, NI Zhi-wei, LIU Xiao
(School of Management, Hefei University of Technology, Hefei 230009, China)

Abstract: Mining of association rules is an important research topic among the various data mining problems. However the common approaches based on support-confidence framework maybe get a great number of redundant and wrong association rules. In order to solve the problems, an interest measure is defined and added to the mining algorithm for association rules. According to the value of interest measure, association rules are classified into positive and negative association rules. The new algorithm can find out the negative-item-contained rules. The experimental result shows that introducing interest measure based on common approach to association rules mining can reduce the scale of positive association rules, and mine a lot of interesting negative association rules.

Key words: association rule; negative association rule; interest measure

0 引 言

关联规则(association rule)是数据挖掘(data mining)研究的主要领域之一,其任务是发现大量数据中项集之间有趣的关联或相关联系。R. Agrawal等人于1993年首先提出关联规则[1]的有关概念,此后许多的学者对关联规则的挖掘问题进行了大量的研究。但通常人们只关注于项集间正关联规则的挖掘,如"买了面包的顾客也可能买牛奶"这样的规则,而忽略了形如"不买咖啡的顾客很可能买牛奶"这样的负规则。在投资分析和竞争分析等许多领域的决策制订过程中,负关联规则的作用不可低估。从系统的完整性角度来看,负关联规则与正关联规则一起为正确决策提供更加全面的信息,正因为如此,负关联规则的研究正受到越来越多的重视。

早在1997年,Brin等人就指出了负规则的重要性[2],Savasere等人阐述了强负关联规则问题[3],Xindong Wu等人在文献[4]中给出了一种PR模型,并且给出了一个能够同时挖掘正、负关联规则的算法。

笔者将兴趣度[5]进行了重新定义,并进一步推广,使其不仅能够适用于负关联规则,而且还能够对关联规则的相关性进行判断,并在此基础上提出一个能同时挖掘正、负关联规则的算法。

1 相关概念

设 $I=\{i_1,i_2,\cdots,i_m\}$ 为项集,包含 k 个项的项集称为 k - 项集,$1\leqslant k\leqslant m$。D 为事物数据库,$D=\{T_1,T_2,\cdots,T_m\}$,其中每个事务 T_i 是项的集合,使得 $T\subseteq I$。关联规则是一个形如 $A\rightarrow B$ 的蕴涵式,其中 $A\subseteq I, B\subseteq I$,且 $A\cap B=\varnothing$。规则 $A\rightarrow B$ 在事务集 D 中成立,且具有支持度 $\mathrm{supp}(A\rightarrow B)$ 和置信度 $\mathrm{conf}(A\rightarrow B)$。这也意味着数据库 D 中有 s 比例的交易 T 中包

收稿日期:2006-08-31
基金项目:安徽省自然科学基金资助项目(050460402)
作者简介:李伟东(1981-),男,山东烟台人,硕士研究生,研究方向为数据挖掘;倪志伟,教授,博士生导师,研究方向为机器学习、数据挖掘。

含 $A \cup B$ 数据项;且数据库 D 中有 c 比例的交易 T 满足"若包含 A 就包含 B 条件"。具体描述为:

$$\text{supp}(A \rightarrow B) = P(A \cup B) \quad (1)$$

$$\text{conf}(A \rightarrow B) = P(B \mid A) \quad (2)$$

满足最小支持度的项集的集合称为频繁项集;满足最小支持度阈值和最小置信度阈值的规则称为强关联规则。

2　问题的提出

目前,关联规则的评价标准主要有两个,即支持度(support)和置信度(confidence),然而,按照现有标准来产生关联规则,会产生大量不相关、甚至是误导的关联规则。现来考察以下的例子。

研究一下关联规则:"买牛奶→买咖啡",它的支持度 Sup=20/100=0.2,置信度 Conf=20/25=0.8。如果设定最小支持度和最小可信度分别为 0.2 和 0.6,那么按照 Agrawal 的定义,该规则会被作为强关联规则挖掘出来。由此得出结论,刺激顾客对牛奶的购买欲望将增加咖啡的销售量。

然而,事实上由表 1 的交易数据库可以看出不买牛奶就会买咖啡(不买牛奶→买咖啡)的可能性更大(70/75=93.3%)。也就是说买牛奶的人会买咖啡的可能性小于不买牛奶的人买咖啡的可能性。这就产生了一个矛盾,到底是买牛奶→买咖啡,还是不买牛奶→买咖啡。

表 1　某交易数据库示例

	买咖啡	不买咖啡	合计
买牛奶	20	5	25
不买牛奶	70	5	75
合计	90	10	100

3　问题的分析及解决方法

从上例中可看出,一条即使可信度和支持度都很高的规则,它的实际价值已经没有人们期望的那么高了,更严重的话,这条规则确实会是误导性的。因此,人们引入了新的标准——兴趣度来加强对强关联规则的判定。

为了解决支持度 - 置信度模型引起的问题,文献[6]提出了改进的兴趣度的标准,文献[7]利用统计性质来定义事务间的兴趣度。另外还有信任度[8]、收集强度[9]等定义。但是上述文献中提到的度量标准的大小仍然无法有效说明规则体对结论的影响程度。换句话说,不能反映出 A 出现的条件下 B 出现的概率 $P(B \mid A)$ 与没有任何前提条件下 B 出现的概率 $P(B)$ 之间产生的差别,即 $P(B \mid A)$ 与 $P(B)$ 之间的差值。若 $P(B \mid A)$ 与 $P(B)$ 的值相差较大,则说明 A 的出现对 B 的影响很大,规则 $A \Rightarrow B$ 是有趣的,它将给用户的决策过程提供有意义的指导信息。下面给出基于概率差值的兴趣度定义。

3.1　基于概率差值的兴趣度

$$\text{Interest}(A \rightarrow B) = P(B \mid A) - P(B) \quad (3)$$

根据式(3)的定义,Interest($A \rightarrow B$) 的度量有两种可能的情况:

(1) 如果 Interest > 0,那么 A 和 B 正相关。

(2) 如果 Interest < 0,那么 A 和 B 负相关。

表 1 中的关联规则 $R1$:"买牛奶(A) → 买咖啡(B)" 其兴趣度为:

$$\text{Interest}(R1) = P(B \mid A) - P(B) = -0.1$$

由此得出,该规则体和结论是负相关的,反映正相关的规则体"买牛奶(A) → 买咖啡(B)" 应该被淘汰。

3.2　正负关联规则间的兴趣度关系

项集 A,B 间有 4 种形式的关联规则:$A \rightarrow B$,$A \rightarrow \neg B$,$\neg A \rightarrow B$ 和 $\neg A \rightarrow \neg B$。它们之间存在下面的内在联系:

定理 1 如果 Interest($A \rightarrow B$) > 0,那么

(1) Interest($\neg A \rightarrow B$) < 0;

(2) Interest($A \rightarrow \neg B$) < 0;

(3) Interest($\neg A \rightarrow \neg B$) > 0。

根据文献[10]的推论 1 及公式(3),给出结论(2)的证明。

证明:Interest($A \rightarrow \neg B$) = $P(\neg B \mid A) - P(\neg B) = 1 - P(B \mid A) - (1 - P(B)) = -P(B \mid A) + P(B) = -(P(B \mid A) - P(B))$

由条件 Interest($A \rightarrow B$) > 0 可知,$-(P(B \mid A) - P(B)) < 0$,即 Interest($A \rightarrow \neg B$) < 0。证毕。

结论(1)和(3)同理可证。

反之亦反之。

定理 1 说明规则 $A \rightarrow B$(或 $\neg A \rightarrow \neg B$) 和 $A \rightarrow \neg B$(或 $\neg A \rightarrow B$) 不会同时作为有效规则,从而有效防止自相矛盾的规则产生。

4　算法设计

根据上面的讨论,给出一个基于概论差的兴趣度的算法,该算法能够判断项集间的相关性并能同时挖掘出频繁项集中的正、负关联规则。在算法中,假定频繁项集 L 已求得。

算法: PN_ RI

输入: L: 频繁项集; min_ conf:最小置信度

输出: PAR:正关联规则集合; NAR:负关联规则

集合

(1)PAR = Ø;NAR = Ø; /* 正、负关联规则集合 */

(2)//从频繁项集 L 产生所有可能的有意义的正、负关联规则

for each itemset I in L {

for each X, Y ($I = X \cup Y$) and $X \cup Y = Ø$ {

①ri = support($X \cup Y$)/support(X) − support(Y)); /* 计算 X、Y 的兴趣度 */

②if (ri > 0) then //正相关

if conf($X \rightarrow Y$) > min_conf then PAR = PAR ∪ {$X \rightarrow Y$}

if conf($\neg X \rightarrow \neg Y$) > min_conf then NAR = NAR ∪ {$\neg X \rightarrow \neg Y$}

③if (ri < 0) then //负相关

if conf($X \rightarrow \neg Y$) > min_conf then NAR = NAR ∪ {$X \rightarrow \neg Y$}

if conf($\neg X \rightarrow Y$) > min_conf then NAR = NAR ∪ {$\neg X \rightarrow Y$}

}

}

(3) return PAR and NAR /* 返回所有的有意义的正、负关联规则,结束算法 */

算法同时生成了频繁项集 L 的正关联规则集(PAR)与负关联规则集(NAR)。第一步将 PAR 和 NAR 初始化为空集;第二步,首先检查是否满足最小兴趣度,然后根据 ri 的值判断相关性,并产生规则,其中,当 ri 正相关时,由步骤 ② 产生形如 $X \rightarrow Y$ 和 $\neg X \rightarrow \neg Y$ 的规则,当 ri 负相关时步骤 ③ 产生形如 $X \rightarrow \neg Y$ 以及 $\neg X \rightarrow Y$ 的规则。第三步返回结果 PAR 和 NAR,结束整个算法。

5 实 验

文中用某商场部分用户的购买记录为例验证 PN_RI 的有效性。实验中采用了 100 位顾客对 15 种商品的购买记录。实验在 P Ⅲ 800, 256RAM, WIN2000,Java 环境下进行。最小置信度(min_conf)设定为 0.6 时,两种算法生成的正关联规则总数随着支持度阈值变化的测试结果见图 1,两种算法生成的关联规则总数随着支持度阈值变化的测试结果见图 2。

从图 1 中可以看到,文中 PN_RI 算法生成的正关联规则数目明显少于经典的 Apriori 算法,这是因为文中的算法中引入了基于概率差的兴趣度,将一些负相关的关联规则给过滤掉了,这就相当于在经典的 Apriori 算法生成的规则中进行了二次过滤,从而使规则变得更加客观、合理。

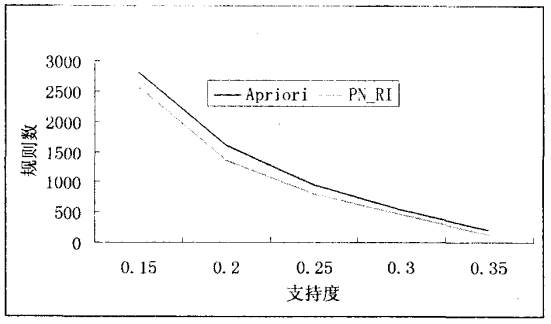

图 1 正关联规则数目比较

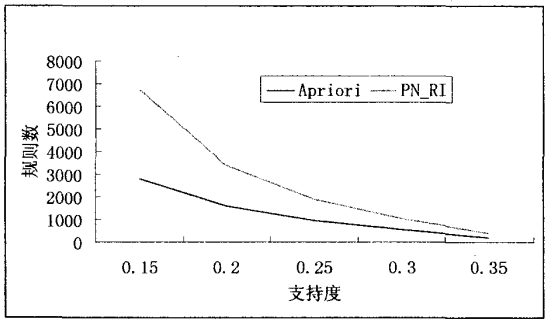

图 2 关联规则总数目比较

然而,从图 2 中可以看到,PN_RI 算法产生的关联规则总数却明显多于经典的 Apriori 算法。这是因为经典的 Apriori 算法只关注于满足最小置信度的强关联规则的挖掘,而忽略了其他一些可能更新奇的、更有价值的负关联规则的挖掘。文中的 PN_RI 算法通过兴趣度的判断,不仅挖掘到一些有效的正关联规则,而且还挖掘到了许多有意义的负关联规则。实验结果说明算法 PN_RI 是有效的。

6 结 论

传统的 Apriori 算法得到的关联规则并不总是相关的、有价值的,有时甚至是误导的。文中引入了基于概率差值的兴趣度,并对其进行了扩展,使其不仅能检测并删除相互矛盾的规则,而且能适用于负关联规则的挖掘。然后在此基础上提出了一个能同时挖掘正负关联规则的算法,实验表明该算法是有效的。

参考文献:

[1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large database[C]// Proceeding of the 1993 ACMSIGMOD International conference on Management of Data. New York: ACM Press, 1993: 207 − 216.

[2] Brin S, Motwani R, Silverstein C. Beyond market Generalizing

(下转第 86 页)

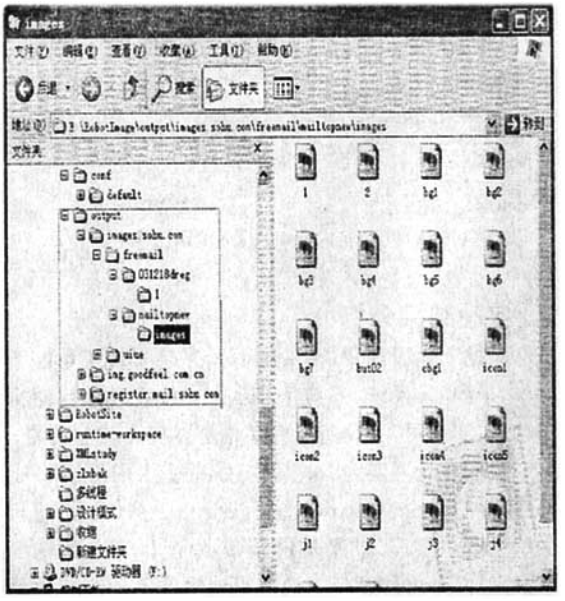

图 6 图像搜索插件工作结果

Robot 工作模式进行灵活的配置和调整，以使信息的采集具有针对性，从而降低搜索服务运营的整体开销。我们所设计的可配置 Robot 系统，在继承常规 Robot 系统的基础上，通过应用面向对象的设计思想，构建了可扩充的服务提供接口，可以使 Robot 根据实际需求进行灵活的控制。实验表明此系统的设计方案可行有效，通过开发所需要的插件和规则，即可以实现不同 Web 资源的搜索，使其灵活应用于各种搜索环境，实现了增量式开发，节省了开发成本。

目前，此可配置 Robot 系统还存在许多需要改进之处，如：缺乏动态性能的监测和调整、未能构建分布式存储机制等，在下一个阶段，将针对这些方面展开工作，使系统得到完善。

参考文献：

[1] Koster M. Robots in the web: threat or treat? [EB/OL]. 1995-04[2006-07-02]. http://www.robotstxt.org./wc/threat-or-treat.html.

[2] Chau M, Chen HsinChum. Personalized and Focused Web Spiders[EB/OL]. 2003[2006-06-20]. http://citeseer.ist.psu.edu/548327.html.

[3] Gray M. Internet Growth and Statistics: Credits and Background[EB/OL]. [2006-06-20]. http://www.mit.edu/-mkgray/net/background.html.

[4] Chau M, Zeng D, Chen Hinchun. Personalized spiders for web search and analysis[C]// Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries. New York, USA: ACM press, 2001: 79-87.

[5] Arasu A, Cho Junghoo, Garcia-Molina H, et al. Searching the Web[J]. ACM Transactions on Internet Technology, 2001(1): 2-43.

[6] Menczer F, Pant G, Srinivasan P. Topical web crawlers: Evaluating adaptive algorithms[J]. ACM Transactions on Internet Technology, 2004(4): 378-419.

[7] 埃克尔. Java 编程思想[M]. 陈昊鹏，等译. 北京：机械工业出版社，2005.

(上接第 82 页)

association rules to correlations[C]// Proceeding of the ACM-SIGMOD Conference 1997. New York: ACM Press, 1997: 265-276.

[3] Savasere A, Omiecinski E, Navathe S. Mining for strong negative associations in a large database of customer transaction[C]// Proceedings of the IEEE 14th International Conference on Data Engineering. Los Alamitos: IEEE-CS, 1998: 494-502.

[4] Wu Xindong, Zhang Chengqi, Zhang Shichao. Mining both positive and negative association rules[C]// Proceedings of the 19th International Conference on Machine Learning(ICML-2002). San Francisco: Morgan Kaufmann Publishers, 2002: 658-665.

[5] 周 欣，沙朝锋，朱央勇，等. 兴趣度——关联规则的又一个阈值[J]. 计算机研究与发展，2000，37(5)：627-633.

[6] Srikant R, Agrawal R. Mining quantitative association rules in large relational tables[C]// Proceedings of the 1996 ACM SIGMOD international conference on Management of Data. Montreal, Canada: [s.n.], 1996: 1-2.

[7] Brin S, Motwani R, Silverstein C. Beyond market baskets: generalizing association rules to correlations[C]// Proceedings of the 1997 ACM SIGMOD international conference on Management of Data. Tucson, USA: [s.n.], 1997: 265-276.

[8] Brin S, Motwani R, Ullman J D, et al. Dynamic itemset counting and implication rules for market basket data[C]// Proceedings of the 1997 ACM SIGMOD international conference on Management of Data. Tucson, USA: [s.n.], 1997: 255-264.

[9] Aggarwal C C, Yu P S. Online generation of association rules, Data Engineering[C]// 1998 Proceedings of the IEEE 14th International conference on Data Engineering. Orlando, Florida, USA: [s.n.], 1998: 402-411.

[10] 董祥军，王舒静，宋瀚涛，等. 负关联规则的研究[J]. 计算机工程与应用，2004，40(11)：978-981.