文章编号: 1001-0920 (2015) 10-1729-13 DOI: 10.13195/j.kzyjc.2014.1102

基于项权值变化和SCCI框架的加权正负关联规则挖掘

黄名选[1], 黄发良[2], 严小卫[3], 兰慧红[4]

(1. 广西财经学院 信息与统计学院，南宁 530003；2. 福建师范大学 软件学院，福州 350007；
3. 美国系统生物研究所，西雅图 WA 98109；4. 广西教育学院 人事处，南宁 530023)

摘 要: 给出项权值变化的数据模型形式化表示, 构建新的加权项集剪枝策略及其模式评价框架SCCI (support-confidence-correlation-interest), 提出基于项权值变化和SCCI评价框架的加权正负关联规则挖掘算法. 该算法考虑了项权值变化的数据特点, 采用新的剪枝方法和评价框架, 通过项集权值简单计算和比较, 挖掘有效的加权正负关联规则. 实验结果表明, 该算法能够有效地减少候选项集数量和挖掘时间, 挖掘出有趣的关联模式, 避免无效模式出现, 挖掘效率高于相比较的现有算法, 解决了项权值变化的加权负模式挖掘问题.

关键词: 数据挖掘；加权关联模式；正负关联规则；频繁项集

中图分类号: TP391 **文献标志码:** A

Weighted positive and negative association rules mining based on dynamic item weight and SCCI framework

HUANG Ming-xuan[1], *HUANG Fa-liang*[2], *YAN Xiao-wei*[3], *LAN Hui-hong*[4]

(1. College of Information and Statistics, Guangxi University of Finance and Economics, Nanning 530003, China; 2. Faculty of Software, Fujian Normal University, Fuzhou 350007, China; 3. Institute for Systems Biology of USA, Seattle WA 98109, USA; 4. Personnel Department, Guangxi College of Education, Nanning 530023, China. Correspondent: HUANG Ming-xuan, E-mail: huangmx@mailbox.gxnu.edu.cn)

Abstract: The formal definition of data model for dynamic item weight is given, and a new pruning strategy for weighted itemsets, as well as an evaluation framework, support-confidence-correlation-interest(SCCI), of weighted association patterns is proposed. Based on dynamic item weight and SCCI, an algorithm for the mining of weighted positive and negative association rules is presented. With the characteristics of the dynamic item weighted data taken into consideration, new pruning methods and evaluation standards are used. Effective weighted frequent itemsets, as well as negative itemsets are mined from the massive weighted database by using the proposed algorithm, and valid weighted positive and negative association rules can be mined by means of simple computation and comparison of itermset weight. The experimental results show that, by using the proposed algorithm, the mining time and the number of candidate itemsets are effectively reduced. Interesting association patterms are obtained, and ineffective patterns are successfully avoided. Compared with the existing mining algorithms, the mining efficiency of this approach is greatly improved, and the problem of the mining of weighted negative patterns is solved based on dynamic item weight.

Keywords: data mining; weighted association pattern; positive and negative association rule; frequent itemset

0 引 言

关联规则挖掘是数据挖掘领域的研究热点之一, 20多年来, 其研究主要集中在基于项频度的挖掘和基于项权值的挖掘两个方面. 基于项频度的挖掘是早期的正负关联规则挖掘技术, 其特点是只考虑项目频度, 平等一致地处理项目. Agrawal等[1]提出的Apriori算法是经典的关联规则挖掘算法. 此后出现了一些改进算法, 有的从剪枝策略方面进行改进, 以提高挖掘效

收稿日期: 2014-07-15；**修回日期:** 2015-01-06.

基金项目: 国家自然科学基金项目(61262028, 61363037)；广西自然科学基金项目(2012GXNSFAA053235)；教育部人文社会研究青年基金项目(12YJCZH074)；广西财经学院数量经济学创新团队项目(2014CX01)；广西教育厅科研项目(201203YB225, 2013LX236, KY2015YB337, KY2015YB483)；广西教育学院科研项目(B2012007).

作者简介: 黄名选(1966−), 男, 教授, 从事数据挖掘和信息检索等研究；黄发良(1975−), 男, 副教授, 博士, 从事数据挖掘和智能计算等研究.

率, 比如Narmadha等[2]提出的一种新的挖掘关联规则的剪枝策略; 有的从关联规则评价方式进行改进, 比如董杰等[3]提出的事务间频繁闭项集的概念及其挖掘算法, GLass[4]提出的两种新的关联规则兴趣度确定方法, 均取得了良好的挖掘效果; 有的改进了挖掘方式, 提出新的挖掘算法, 比如何波[5]提出的一种基于频繁模式树的分布式关联规则挖掘算法, 大幅度减少了候选项集, Shaheen等[6]提出的基于上下文的时空关联规则挖掘算法, 用于挖掘时空数据中的关联规则, 这些成果在挖掘性能上都有良好的表现. 随着研究的深入, 基于项频度的负关联规则挖掘技术得到了广泛研究, 其中最具代表性的算法是Wu等[7]提出的正负关联规则挖掘算法. 在此基础上, 一种能按层次设置而产生的负关联规则挖掘算法[8], 基于多支持度的正负关联规则挖掘算法[9], 基于X^2检验的正负属性关联规则挖掘算法[10], 以及基于频繁模式树的正负关联规则挖掘算法[11]等相继被提出, 这些算法都是当前基于项频度挖掘研究中的最新成果.

基于项频度挖掘的缺陷是: 只考虑项频度, 没有考虑存在项目权值的情况. 针对上述问题, 基于项权值的挖掘得到了深入讨论和研究, 其特点是引入项权值, 以体现项目之间具有不同的重要性以及项目在事务记录中具有不同的权值. 根据项权值的来源不同, 基于项权值的挖掘可分为基于项权值固定的关联规则挖掘和基于项权值变化的关联规则挖掘.

基于项权值固定的挖掘特点是: 项权值来源于用户或领域专家的主观设置, 独立于事务, 并且固定不变. 其典型算法是Cai等[12]于1998年提出的加权关联规则挖掘算法MINWAL. 在此基础上, Yun等[13]提出了近似加权频繁模式概念及其挖掘算法, 在噪音环境下能稳定地挖掘出加权频繁项集; Pears等[14]提出了一种基于粒子群优化的加权关联规则挖掘算法, 该算法使用粒子群优化技术来分配项集的权值, 取得了良好的挖掘效果. 近年来, 具有反单调性的加权支持度框架及其加权关联规则挖掘算法[15], 基于加权频繁项集树的加权频繁项集挖掘算法[16], 以及基于连通模型图的加权关联规则挖掘算法[17]等均是当前基于项权值固定的挖掘研究的最新成果, 其在挖掘效率和挖掘性能方面都有良好表现. 2008年以来, 基于项权值固定的加权负关联规则挖掘得到了重视和研究, 其典型算法是Jiang等[18]提出的WNAILMS算法以及Zhao等[19]提出的WNRIF算法, 解决了基于项权值固定的加权负关联规则挖掘问题.

基于项权值固定的挖掘存在的不足是: 只考虑项目之间的重要性, 没有考虑项目在各个事务记录中存在不同权值(即项权值变化)的情况. 基于项权值变化的挖掘能够有效地解决上述问题, 其特点是项权值依赖于事务, 并随事务记录变化而变化, 其典型算法是谭义红等[20]提出的KWEstimate算法及其一些改进算法, 例如矩阵加权关联规则挖掘算法MWARM[21], 解决了基于项权值变化的加权正关联规则挖掘问题.

当前, 基于项频度挖掘和基于项权值固定的挖掘研究比较充分, 而基于项权值变化的关联模式挖掘的研究并不深入, 其国内外研究报道不多. 目前普遍采用基于频度的挖掘方法来处理具有项权值变化特征的数据, 导致大量虚假的、冗余的和无趣的关联模式产生. 随着信息技术和网络的发展, 具有项权值变化特征的数据迅猛增加, 已经成为海量大数据, 例如, Web文本信息数据、教育系统的课程考试成绩数据等, 很有必要发展和研究基于项权值变化的挖掘技术. 鉴于此, 本文对基于项权值变化的正负关联规则挖掘技术进行深入探讨和研究, 在深入分析项权值变化的加权数据模型特点基础上, 给出项权值固定和项权值变化的数据模型形式化表示, 构建一种新的加权项集剪枝策略和基于项权值变化的关联模式评价框架SCCI, 提出一种新的基于项权值变化和SCCI的加权正负关联规则挖掘算法——WPNARM-SCCI. 该算法考虑了项权值依赖于事务的特点, 采用新的剪枝方法和模式评价标准, 挖掘有趣的加权频繁项集和负项集, 通过项集权值的简单计算和比较, 从频繁项集和负项集中挖掘出有效的加权正负关联规则. 以中英文本标准测试集(CWT 200 g和NTCIR-5)语料为实验数据的实验结果表明, WPNARM-SCCI算法具有良好的剪枝效果和扩展性, 与现有基于频度的挖掘算法和加权模式挖掘算法相比较, 其挖掘时间和关联模式数量明显减少, 能挖掘出有趣的正负关联模式, 避免了无效的和无趣的模式出现, 挖掘效率得到极大提高, 解决了基于项权值变化的加权负模式挖掘问题.

1 基本概念及相关定理

1.1 数据模型形式化表示及其主要区别

项权值固定的数据模型(FWDM)如表1所示. 其中: “1”表示项目在事务中出现, “0”表示不出现的情况, 其模型可以形式化为如下的三元组表示:

$$\mathrm{FWDM} = (T, I, W_I).$$

其中: $T = \{T_1, T_2, \cdots, T_n\}$是所有事务的有限集合; $I = \{i_1, i_2, \cdots, i_m\}$是$T$中全部项目的有限集合; W_I是I上对应项目权值的项目权重关系集合, 即

$$W_I = \{\langle i_1, w_1\rangle, \langle i_2, w_2\rangle, \cdots, \langle i_m, w_m\rangle\}.$$

项权值变化的数据模型(DWDM)如表2所示, 其形式化为如下的三元组表示:

$$\mathrm{DWDM} = (T, I, W_{IT}).$$

其中: T 和 I 与上述FWDM相同; W_{IT} 是 I 在事务 T 上对应项目权值的项目权重关系集合, 即

$$W_{IT}=\{\langle i_1,T_1,w_{11}\rangle,\langle i_1,T_2,w_{12}\rangle,\cdots,\langle i_1,T_n,w_{1n}\rangle,\langle i_2,T_1,w_{21}\rangle,\langle i_2,T_2,w_{22}\rangle,\cdots,\langle i_2,T_n,w_{2n}\rangle,\cdots,\langle i_m,T_1,w_{m1}\rangle,\langle i_m,T_2,w_{m2}\rangle,\cdots,\langle i_m,T_n,w_{mn}\rangle\}.$$

表 1　项权值固定的数据模型 (FWDM)

事务	$i_1:w_1$	$i_2:w_2$	$\cdots$	$i_m:w_m$
T_1	1/0	1/0	$\cdots$	1/0
T_2	1/0	1/0	$\cdots$	1/0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
T_n	1/0	1/0	$\cdots$	1/0

表 2　项权值变化的数据模型 (DWDM)

事务	i_1	i_2	$\cdots$	i_m
T_1	w_{11}	w_{21}	$\cdots$	w_{m1}
T_2	w_{12}	w_{22}	$\cdots$	w_{m2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
T_n	w_{1n}	w_{2n}	$\cdots$	w_{mn}

FWDM和DWDM都引入了项目权值, 从而体现了项目之间具有不同的重要性. 但是, 它们却是两种不同的数据类型, 主要区别如下:

1) 项目权值来源和设置方式不同. FWDM的项目权值主要由用户或领域专家根据项目重要性不同而设置; DWDM的项目权值并不是主观设置, 而是客观存在于事务记录中, 体现同一项目在不同事务记录中具有不同的重要性, 其权值计算是根据具体数据的项目权值计算方法而得到的. 例如TF-IDF (term frequency-inverse document frequency), 其权值计算方法是目前文本数据特征词项目权值计算最常见的方法[22].

2) 项目权值与事务数据库的关系不同. FWDM的项目权值独立于事务数据库, 一旦设置后便固定不变; 而DWDM的项目权值则与事务数据库关系密切, 依赖于具体的事务, 随事务记录不同而变化.

FWDM和DWDM的不同特点决定了基于项权值固定的模式挖掘与基于项权值变化的模式挖掘具有本质的区别. 首先, 其模式支持度计算方法不同, 前者的支持度计算只针对项权值固定不变的情况, 而后者的支持度计算则必须考虑项权值随着事务的不同而变化的特点; 其次, 其挖掘的数据对象不同, 前者的挖掘方法适用于具有项权值固定特点的超市交易数据, 并不适合项权值变化的数据, 而后者的挖掘方法则主要针对具有DWDM特点的数据, 如文本信息数据, 教育系统中的教务数据等.

1.2　加权支持度与置信度

基于权值变化的加权项集支持度(dynamic weighted support, dwsup)是项集 I 的项目权值平均值与该项集出现的概率(即无加权支持度)的乘积, $\mathrm{dwsup}(I)$ 的计算公式[20-21]如下:

$$\mathrm{dwsup}(I)=\mathrm{Iwa}(I)\times\sup(I)=\frac{w_I}{n_I\times k_I}\times\sup(I)=\frac{w_I}{n_I\times k_I}. \tag{1}$$

其中: $w_I=\sum_{T_b\in(T)}\sum_{i_a\in(I)}w_{ab}$表示项集 I 在所有事务 T 中的权值总和, $w_{ab}(1\leqslant a\leqslant m,1\leqslant b\leqslant n)$ 为项集 I 中的全部项目在事务 T_b 出现时项目 i_a 在事务 T_b 的权值; n 为 T 中事务记录总数, n_I 为项集 I 中全部项目同时出现在事务记录的次数, k_I 为项集 I 的项目个数(即项集长度); $\sup(I)$ 为项集 I 在事务 T 中的无加权支持度, 其计算公式[1]如下:

$$\sup(I)=\frac{w_I}{n}; \tag{2}$$

$\mathrm{Iwa}(I)$ 为加权项集 I 的项目权值平均值 (average value of itemset weight, Iwa), 其计算公式如下:

$$\mathrm{Iwa}(I)=\frac{w_I}{n_I\times k_I}. \tag{3}$$

根据概率性质, 加权负项集支持度的计算公式如下:

$$\mathrm{dwsup}(\neg I)=1-\mathrm{dwsup}(I), \tag{4}$$

$$\mathrm{dwsup}(I_1,\neg I_2)=\mathrm{dwsup}(I_1)-\mathrm{dwsup}(I_1,I_2), \tag{5}$$

$$\mathrm{dwsup}(\neg I_1,I_2)=\mathrm{dwsup}(I_2)-\mathrm{dwsup}(I_1,I_2), \tag{6}$$

$$\mathrm{dwsup}(\neg I_1,\neg I_2)=1-\mathrm{dwsup}(I_1)-\mathrm{dwsup}(I_2)+\mathrm{dwsup}(I_1,I_2). \tag{7}$$

其中: 项集 (I_1,I_2) 的频度 $n(I_1,I_2)$ 是指其子项集 I_1 和 I_2 在同一事务记录中同时出现的次数; 负项集 $(I_1,\neg I_2)$ 的频度 $n(I_1,\neg I_2)$是指其子项集 I_1在事务记录中出现, 同时 I_2不在该事务记录中出现的次数; $n(\neg I_1,I_2)$ 与 $n(I_1,\neg I_2)$ 类似, $n(\neg I_1,\neg I_2)$ 是指其子项集 I_1 和 I_2 不同时出现在同一事务记录中的次数.

定义 1 (加权频繁项集和负项集)　设最小支持度阈值为ms, 在基于项权值变化的挖掘中, 若

$$\mathrm{dwsup}(I)\geqslant\mathrm{ms},$$

则称项集 I 为加权频繁项集; 当加权项集 I_1 和 I_2 是频繁项集时, 若 $\mathrm{dwsup}(I_1,I_2)<\mathrm{ms}$, 则称项集 (I_1,I_2) 为加权负项集, 其有3种形式, 即 $(I_1,\neg I_2)$、$(\neg I_1,I_2)$ 和 $(\neg I_1,\neg I_2)$.

定义 2 (加权正负关联规则置信度)　基于项权值变化的加权正负关联规则置信度(dynamic weighted confidence, dwconf)计算公式如下:

$$\mathrm{dwconf}(I_1\to I_2)=\frac{\mathrm{dwsup}(I_1,I_2)}{\mathrm{dwsup}(I_1)}, \tag{8}$$

$$\mathrm{dwconf}(I_1 \to \neg I_2) = \frac{\mathrm{dwsup}(I_1, \neg I_2)}{\mathrm{dwsup}(I_1)}, \tag{9}$$

$$\mathrm{dwconf}(\neg I_1 \to I_2) = \frac{\mathrm{dwsup}(I_2) - \mathrm{dwsup}(I_1, I_2)}{1 - \mathrm{dwsup}(I_1)}, \tag{10}$$

$$\mathrm{dwconf}(\neg I_1 \to \neg I_2) = \frac{\mathrm{dwsup}(\neg I_1, \neg I_2)}{\mathrm{dwsup}(\neg I_1)}. \tag{11}$$

定义 3 (加权强正负关联规则) 在基于权值变化的挖掘中, 如果 I_1、I_2 和 (I_1, I_2) 是频繁项集, 并且 $\mathrm{dwconf}(I_1 \to I_2) \geqslant \mathrm{mc}$ 和 $\mathrm{dwconf}(\neg I_1 \to \neg I_2) \geqslant \mathrm{mc}$, 或者, 如果 (I_1, I_2) 是负项集, 并且 $\mathrm{dwconf}(\neg I_1 \to \neg I_2) \geqslant \mathrm{mc}$、$\mathrm{dwconf}(I_1 \to \neg I_2) \geqslant \mathrm{mc}$ 和 $\mathrm{dwconf}(\neg I_1 \to I_2) \geqslant \mathrm{mc}$, 则称关联规则 $I_1 \to I_2$ 为加权强正关联规则, $\neg I_1 \to \neg I_2$、$I_1 \to \neg I_2$ 和 $\neg I_1 \to I_2$ 为加权强负关联规则.

1.3 加权项集相关性

基于权值变化的加权项集相关性 (dwCorr) 是指加权项集 (I_1, I_2) 的子项集 I_1 与 I_2 之间 $(I_1 \bigcap I_2 = \varnothing)$ 固有的相关密切程度. 基于无加权项集相关性的定义[8], 给出项权值变化的加权项集相关性计算公式如下:

$$\mathrm{dwCorr}(I_1, I_2) = \frac{\mathrm{dwsup}(I_1, I_2)}{\mathrm{dwsup}(I_1)\mathrm{dwsup}(I_2)} = \frac{\mathrm{Iwa}(I_1, I_2)}{\mathrm{Iwa}(I_1)\mathrm{Iwa}(I_2)} \times \mathrm{Corr}(I_1, I_2), \tag{12}$$

其中 $\mathrm{Corr}(I_1, I_2) = \sup(I_1, I_2)/(\sup(I_1) \times \sup(I_2))$ 是无加权项集 (I_1, I_2) 的相关性[8]. 根据相关性的性质, 项权值变化的加权项集 (I_1, I_2) 相关性具有如下性质:

性质 1 $\mathrm{dwCorr}(I_1, I_2) > 1 \Leftrightarrow$ 加权项集 I_1 与 I_2 成正相关;

性质 2 $\mathrm{dwCorr}(I_1, I_2) < 1 \Leftrightarrow$ 加权项集 I_1 与 I_2 成负相关;

性质 3 $\mathrm{dwCorr}(I_1, I_2) = 1 \Leftrightarrow$ 加权项集 I_1 与 I_2 相互独立, 无相关性.

定理 1 若 $\mathrm{dwCorr}(I_1, I_2) > 1 \Leftrightarrow (\mathrm{dwCorr}(I_1, \neg I_2) < 1, \mathrm{dwISCorr}(\neg I_1, I_2) < 1, \mathrm{dwCorr}(\neg I_1, \neg I_2) > 1)$; 反之, 若 $\mathrm{dwCorr}(I_1, I_2) < 1 \Leftrightarrow (\mathrm{dwCorr}(I_1, \neg I_2) > 1, \mathrm{dwCorr}(\neg I_1, I_2) > 1, \mathrm{dwCorr}(\neg I_1, \neg I_2) < 1)$.

证明 这里仅证明 $\mathrm{dwCorr}(I_1, I_2) > 1$ 的情况.

1) 证定理 1 中“$\Rightarrow$”方向的命题.

因为

$$\mathrm{dwCorr}(I_1, I_2) > 1 \Rightarrow \frac{\mathrm{dwsup}(I_1, I_2)}{\mathrm{dwsup}(I_1)\mathrm{dwsup}(I_2)} > 1 \Rightarrow$$

$$\mathrm{dwsup}(I_1, I_2) > \mathrm{dwsup}(I_1)\mathrm{dwsup}(I_2) \Rightarrow \tag{13}$$

$$\frac{w_{(I_1, I_2)}}{k_{(I_1, I_2)}} > \frac{w_{I_1} \times w_{I_2}}{n \times k_1 \times k_2}, \tag{14}$$

$$\mathrm{dwCorr}(I_1, \neg I_2) = \frac{\mathrm{dwsup}(I_1, \neg I_2)}{\mathrm{dwsup}(I_1)\mathrm{dwsup}(\neg I_2)} \Rightarrow$$

$$\mathrm{dwCorr}(I_1, \neg I_2) = \frac{\mathrm{dwsup}(I_1) - \mathrm{dwsup}(I_1, I_2)}{\mathrm{dwsup}(I_1) - \mathrm{dwsup}(I_1)\mathrm{dwsup}(I_2)}, \tag{15}$$

由 $\mathrm{dwsup}(I_1) > 0, \mathrm{dwsup}(I_2) > 0, \mathrm{dwsup}(I_1, I_2) > 0$ 以及式 (13) 和 (15) 可得

$$\frac{\mathrm{dwsup}(I_1) - \mathrm{dwsup}(I_1, I_2)}{\mathrm{dwsup}(I_1) - \mathrm{dwsup}(I_1)\mathrm{dwsup}(I_2)} < 1 \Rightarrow \mathrm{dwCorr}(I_1, \neg I_2) < 1.$$

所以

$$\mathrm{dwCorr}(I_1, I_2) > 1 \Rightarrow \mathrm{dwCorr}(I_1, \neg I_2) < 1.$$

同理可证

$$\mathrm{dwCorr}(I_1, I_2) > 1 \Rightarrow \mathrm{dwCorr}(\neg I_1, I_2) < 1.$$

又因为

$$\mathrm{dwCorr}(\neg I_1, \neg I_2) = \frac{\mathrm{dwsup}(\neg I_1, \neg I_2)}{\mathrm{dwsup}(\neg I_1)\mathrm{dwsup}(\neg I_2)} \Rightarrow$$

$$\mathrm{dwCorr}(\neg I_1, \neg I_2) = \frac{1 - \mathrm{dwsup}(I_1) - \mathrm{dwsup}(I_2) + \mathrm{dwsup}(I_1, I_2)}{(1 - \mathrm{dwsup}(I_1))(1 - \mathrm{dwsup}(I_2))}, \tag{16}$$

将式 (1) 代入 (16) 可得

$$\mathrm{dwCorr}(\neg I_1, \neg I_2) = \frac{1}{\eta + \dfrac{w_{I_1} \times w_{I_2}}{n \times k_1 \times k_2}} \times \left(\eta + \frac{w_{(I_1, I_2)}}{k_{(I_1, I_2)}}\right), \tag{17}$$

其中 $\eta = n - \dfrac{w_{I_1}}{k_1} - \dfrac{w_{I_2}}{k_2}$.

又因为 $k_1 \geqslant 1, k_2 \geqslant 1, n$ 是事务总数, 相对于 k_1、k_2、w_{I_1} 和 w_{I_2}, 其值是比较大的正数, 故 $\eta \geqslant 0$, 由式 (14) 和 (17) 可得 $\mathrm{dwCorr}(\neg I_1, \neg I_2) > 1$, 所以

$$\mathrm{dwCorr}(I_1, I_2) > 1 \Rightarrow \mathrm{dwCorr}(\neg I_1, \neg I_2) > 1.$$

2) 证定理 1 中“$\Leftarrow$”方向的命题.

因为

$$\mathrm{dwCorr}(I_1, \neg I_2) < 1 \Rightarrow \frac{\mathrm{dwsup}(I_1) - \mathrm{dwsup}(I_1, I_2)}{\mathrm{dwsup}(I_1) - \mathrm{dwsup}(I_1)\mathrm{dwsup}(I_2)} < 1 \Rightarrow \mathrm{dwsup}(I_1, I_2) > \mathrm{dwsup}(I_1)\mathrm{dwsup}(I_2) \Rightarrow \frac{\mathrm{dwsup}(I_1, I_2)}{\mathrm{dwsup}(I_1)\mathrm{dwsup}(I_2)} > 1 \Rightarrow \mathrm{dwCorr}(I_1, I_2) > 1,$$

所以

$$\mathrm{dwCorr}(I_1, \neg I_2) < 1 \Rightarrow \mathrm{dwCorr}(I_1, I_2) > 1.$$

同理可证

$$\mathrm{dwCorr}(\neg I_1, I_2) < 1 \Rightarrow \mathrm{dwCorr}(I_1, I_2) > 1,$$

$$\mathrm{dwCorr}(\neg I_1, \neg I_2) > 1 \Rightarrow \mathrm{dwCorr}(I_1, I_2) > 1.$$

对于 $\mathrm{dwCorr}(I_1, I_2) < 1$ 的情况, 其证明类似于 $\mathrm{dwCorr}(I_1, I_2) > 1$ 的证明过程, 限于篇幅, 证明过程

略. □

定理 1 表明, 若加权项集 $\mathrm{dwCorr}(I_1, I_2) > 1$, 则可以挖掘出加权关联规则 $I_1 \to I_2$ 和 $\neg I_1 \to \neg I_2$ 模式; 当 $\mathrm{dwCorr}(I_1, I_2) < 1$ 时, 可以挖掘出 $I_1 \to \neg I_2$ 和 $I_1 \to \neg I_2$ 模式.

设加权项集 (I_1, I_2) 及其子项集 I_1 和 I_2 的项目个数分别为 k_{12}、k_1 和 k_2, 其在事务 T 中的权值总和分别为 w_{12}、w_1 和 w_2, 参数 $\gamma = k_{12}/(n \times k_1 \times k_2)$, 根据定理 1 和加权相关性性质, 得出如下推论:

推论 1　已知加权项集 $I = (I_1, I_2)$, 且 $I_1 \bigcap I_2 = \varnothing$, 若 $w_{12} > (\gamma \times w_1 \times w_2)$, 则加权子项集 I_1 与 I_2 成正相关, 能挖掘出加权正关联规则 $I_1 \to I_2$ 和负关联规则 $\neg I_1 \to \neg I_2$ 模式.

证明　由式 (1) 和 (12)可以得到

$$\mathrm{dwCorr}(I_1, I_2) = \frac{n \times k_1 \times k_2 \times w_{12}}{k_{12} \times w_1 \times w_2} = \frac{w_{12}}{\dfrac{k_{12}}{n \times k_1 \times k_2} \times w_1 \times w_2} = \frac{w_{12}}{\gamma \times w_1 \times w_2}, \tag{18}$$

由已知 $w_{12} > (\gamma \times w_1 \times w_2)$, 可得 $\mathrm{dwCorr}(I_1, I_2) > 1$, 由性质 1 和定理 1 可知推论 1 成立. □

推论 2　已知加权项集 $I = (I_1, I_2)$, 且 $I_1 \bigcap I_2 = \varnothing$, 若 $w_{12} < (\gamma \times w_1 \times w_2)$, 则加权项集 I_1 与 I_2 成负相关, 能挖掘出加权负关联规则 $I_1 \to \neg I_2$ 和 $\neg I_1 \to I_2$ 模式.

证明　由已知 $w_{12} < (\gamma \times w_1 \times w_2)$ 和式 (18), 可得 $\mathrm{dwCorr}(I_1, I_2) < 1$, 由性质 2 和定理 1 可知推论 2 成立. □

根据推论 1 和推论 2, 只需对项集权值进行简单计算和比较就可以得出项集间是否正负相关, 不需要计算项集间的相关性.

2　基于项权值变化和SCCI框架的加权正负关联规则挖掘

2.1　SCCI: 加权关联模式的评价框架

经典的关联模式评价框架是支持度 (support)-置信度 (confidence), 简称 SC 评价框架. SC 评价框架一般用于评价正关联模式, 其缺点是容易导致冗余的、无趣的和相互矛盾的关联模式出现. 常用的正负关联模式评价标准是支持度 (support)-置信度 (confidence)-相关性 (correlation) 框架, 简称 SCC 评价框架. SCC 评价框架能够很好地避免相互矛盾的关联模型出现, 但也会导致无效的、无趣的模式产生. 针对上述问题, 在基于权值变化的加权数据挖掘环境中, 构建支持度 (support)-置信度 (confidence)-相关性 (correlation)-兴趣度(interest)评价框架, 简称 SCCI 评价框架. 在 SCCI 评价框架中, 根据相关性的值, 将同时满足支持度、置信度和兴趣度要求的关联规则称为有效的加权正负关联模式. SCCI 评价框架使得每个加权关联模式都得到支持度、置信度、相关性和兴趣度的综合评价, 避免了无效的和无趣的关联模式出现, 取得了良好的挖掘效果.

2.2　加权项集剪枝策略

在数据挖掘中, 项集数量呈指数增长, 加权频繁项集和负项集的数量很多, 导致无趣的和无效的项集数量增多. 针对该问题, 兴趣度作为评价关联模式有趣性和新颖性的一个新度量而得到广泛应用[23-24]. 当前一种比较典型的关联模式兴趣度模型是基于概率相关性的兴趣度模型[24], 它反映了项集中两个子项集之间的关系和密切程度. 将基于概率相关性的兴趣度模型引入基于权值变化的加权关联模式评价, 给出如下基于项权值变化的加权项集兴趣度 (dwII) 的计算公式:

$$\mathrm{dwII}(I_1, I_2) = \mathrm{dwsup}(I_1, I_2) - \mathrm{dwsup}(I_1) \times \mathrm{dwsup}(I_2). \tag{19}$$

当 $\mathrm{dwII}(I_1, I_2) = 0$ 时, 表明加权项集 I_1 与 I_2 无相关, 其模式无任何意义, 不会被用户关注; $\mathrm{dwII}(I_1, I_2)$ 越大于 0 或者越小于 0 时, 说明该模式越有趣, 越被用户所关注.

设最小兴趣度阈值为 mi, 给出有趣的加权频繁项集和负项集 (IIs) 判断条件如下:

$$\mathrm{IIs}(I) = \exists I_I, I_2 \subset I : (I_1 \bigcap I_2) = \varnothing \bigwedge (I_1 \bigcup I_2) = I \bigwedge |\mathrm{dwII}(I_1, I_2)| \geqslant \mathrm{mi}. \tag{20}$$

其中: I 为频繁项集或负项集, $\mathrm{IIs}(I)$ 的值是逻辑值.

将满足 $\mathrm{IIs}(I)$ 条件的频繁项集和负项集称为有趣的加权频繁项集和负项集.

加权项集的剪枝策略是: 剪除不满足 $\mathrm{IIs}(I)$ 条件的频繁项集和负项集.

2.3　基于项权值变化和SCCI框架的加权正负关联规则挖掘算法

根据上述定理和推论, 基于项权值变化和 SCCI 框架的加权正负关联规则挖掘的基本思想如下.

1) 首先对项权值变化的加权数据进行预处理, 构建加权数据库 (WD) 和项目库, 从项目库中提取加权候选 1_项集 C_1, 挖掘加权频繁 1_项集 L_1;

2) 从 i-项集 ($i \geqslant 2$)起, 重复进行以下操作, 直到其 L_{i-1}为空: 通过 L_{i-1}产生加权候选 i_项集 C_i, 从 C_i 中挖掘加权频繁 i_项集 L_i 和负 i_项集 N_i;

3) 根据上述 2.2 节的项集剪枝策略, 对频繁项集和负项集进行剪枝, 提取有趣的 L_i 和 N_i;

4) 对于有趣的加权频繁项集和负项集, 根据推论

1 和推论 2 对项集权值进行简单计算比较, 挖掘有效的加权正负关联规则模式.

上述挖掘思想形式化为如下 WPNARM-SCCI (weighted positive and negative association rules minig based on support-confidence-correlation-interest) 算法.

算法 1 WPNARM-SCCI.

输入: 加权数据库 WD, ms, mc, mi;

输出: dwPAR 和 dwNAR (加权强正负关联规则集合).

Step 1: let dwPIS $\leftarrow \varnothing$; dwNIS $\leftarrow \varnothing$; dwPAR $\leftarrow \varnothing$; dwNAR $\leftarrow \varnothing$; // dwPIS 为加权频繁项集集合, dwNIS 为加权负项集集合.

Step 2: let $(C_1, C_1.\ \text{weight}, C_1.\ \text{Count}) \leftarrow$ scan C_1(WD).

Step 3: for each itemset c in C_1 do

begin // $|D|$ 表示数据库 (WD) 总记录数.

Step 3.1: let $L_1 \leftarrow \{c|c \in C_1 \bigwedge(\text{dwsup}(c) = (c.\text{weight}/|D|) >= \text{ms})\}$;

Step 3.2: let dwPIS $\leftarrow$ dwPIS $\bigcup L_1$;

end.

Step 4: for $(i = 2; L_i \neq \varnothing; i++)$ do

begin // 生成所有可能的加权正负 i_项集

Step 4.1: let $C_i \leftarrow \text{AprioriJoin}(L_{i-1})$;

Step 4.2: let (C_i. weight, C_i. Count) $\leftarrow$ scan C_i(WD);

Step 4.3: for each itemset c in C_i do

$\{L_i \leftarrow \{c|\text{dwsup}(c) = (c.\text{weight}/|D| \times i) >= \text{ms}\}$;

$N_i \leftarrow C_i - L_i$; dwPIS $\leftarrow$ dwPIS $\bigcup L_i$;

dwNIS $\leftarrow$ dwNIS $\bigcup N_i$; $\}$

end.

Step 5: for each frequent itemset L_i in dwPIS do

begin // 频繁项集剪枝

Step 5.1: if $(\exists I_1, I_2 \subset L_i : (I_1 \bigcap I_2) = \varnothing \bigwedge (I_1 \bigcup I_2) = L_i \bigwedge((|\text{dwII}(I_1, I_2) = \text{dwsup}(I_1, I_2) - \text{dwsup}(I_1) \times \text{dwsup}(I_2)|) \geqslant \text{mi}))$ then

$\text{IIs}(L_i) = \text{true}$ else $\text{IIs}(L_i) = \text{false}$;

Step 5.2: if $\text{IIs}(L_i) = \text{false}$ then dwPIS $\leftarrow$ dwPIS $- L_i$;

end.

Step 6: for each negative itemset N_i in dwNIS do

begin // 负项集剪枝

Step 6.1: if $(\exists I_1, I_2 \subset N_i : (I_1 \bigcap I_2) = \varnothing \bigwedge (I_1 \bigcup I_2) = N_i \bigwedge((|\text{dwII}(I_1, I_2) = \text{dwsup}(I_1, I_2) - \text{dwsup}(I_1) \times \text{dwsup}(I_2)|) \geqslant \text{mi}))$ then

$\text{IIs}(N_i) = \text{true}$ else $\text{IIs}(N_i) = \text{false}$;

Step 6.2: if $\text{IIs}(N_i) = \text{false}$ then dwNIS $\leftarrow$ dwNIS $- N_i$;

end.

Step 7: for each frequent itemset L_i in dwPIS do

for each expression $I_1 \bigcap I_2 = \varnothing$ and $I_1 \bigcup I_2 = I$ and $\text{dwsup}(I_1) \geqslant \text{ms}$ and $\text{dwsup}(I_2) \geqslant \text{ms}$ do

begin

Step 7.1: $\gamma = k_{12}/(|D| \times k_1 \times k_2)$;

Step 7.2: if $w_{12} > (\gamma \times w_1 \times w_2)$ then

begin

if $\text{dwsup}(I_1, I_2) \geqslant \text{ms}$ then {if $\text{dwconf}(I_1 \rightarrow I_2) \geqslant \text{mc}$, then dwPAR $\leftarrow$ dwPAR $\bigcup\{I_1 \rightarrow I_2\}$; if $\text{dwconf}(I_2 \rightarrow I_1) \geqslant \text{mc}$, then dwPAR $\leftarrow$ dwPAR $\bigcup\{I_2 \rightarrow I_1\}$;}

if $\text{dwsup}(\neg I_1, \neg I_2) \geqslant$ ms then {if $\text{dwconf}(\neg I_1 \rightarrow \neg I_2) \geqslant$ mc then dwNAR $\leftarrow$ dwNAR $\bigcup\{\neg I_1 \rightarrow \neg I_2\}$; if $\text{dwconf}(\neg I_2 \rightarrow \neg I_1) \geqslant$ mc then dwNAR $\leftarrow$ dwNAR $\bigcup\{\neg I_2 \rightarrow \neg I_1\}$;}

end;

Step 7.3: if $w_{12} < (\gamma \times w_1 \times w_2)$ then

begin

if $\text{dwsup}(I_1, \neg I_2) \geqslant \text{ms}$ then {if $\text{dwconf}(I_1 \rightarrow \neg I_2) \geqslant \text{mc}$ then dwNAR $\leftarrow$ dwNAR $\bigcup\{I_1 \rightarrow \neg I_2\}$; if $\text{dwconf}(\neg I_2 \rightarrow I_1) \geqslant$ mc then dwNAR $\leftarrow$ dwNAR $\bigcup\{\neg I_2 \rightarrow I_1\}$;}

if $\text{dwsup}(\neg I_1, I_2) \geqslant$ ms then {if $\text{dwconf}(\neg I_1 \rightarrow I_2) \geqslant$ mc then dwNAR $\Leftarrow$ dwNAR $\bigcup\{\neg I_1 \rightarrow I_2\}$; if $\text{dwconf}(\neg I_2 \rightarrow I_1) \geqslant$ mc then dwNAR $\leftarrow$ dwNAR $\bigcup\{\neg I_2 \rightarrow I_1\}$;}

end;

end.

Step 8: for each negative itemset N_i in dwNIS do

for each expression $I_1 \bigcap I_2 = \varnothing$ and $I_1 \bigcup I_2 = N_i$ and $\text{dwsup}(I_1) \geqslant \text{ms}$ and $\text{dwsup}(I_2) \geqslant \text{ms}$ do

begin

Step 8.1: $\gamma = k_{12}/(|D| \times k_1 \times k_2)$

Step 8.2: if $w_{12} > (\gamma \times w_1 \times w_2)$ and $\text{dwsup}(\neg I_1,$

$\neg I_2) \geqslant$ ms then {if dwconf($\neg I_1 \rightarrow \neg I_2) \geqslant$ mc then dwNAR $\leftarrow$ dwNAR$\bigcup\{\neg I_1 \rightarrow \neg I_2\}$; if dwconf($\neg I_2 \rightarrow \neg I_1) \geqslant$ mc then dwNAR $\leftarrow$ dwNAR$\bigcup\{\neg I_2 \rightarrow \neg I_1\};\}$

Step 8.3: if $w_{12} < (\gamma \times w_1 \times w_2)$ then

begin

if dwsup$(I_1, \neg I_2) \geqslant$ ms then {if dwconf$(I_1 \rightarrow \neg I_2) \geqslant$ mc, then dwNAR $\leftarrow$ dwNAR$\bigcup\{I_1 \rightarrow \neg I_2\}$; if dwconf$(\neg I_2 \rightarrow I_1) \geqslant$ mc then dwNAR $\leftarrow$ dwNAR$\bigcup\{\neg I_2 \rightarrow I_1\};\}$

if dwsup$(\neg I_1, I_2) \geqslant$ ms then {if dwconf$(\neg I_1 \rightarrow I_2) \geqslant$ mc then dwNAR $\leftarrow$ dwNAR$\bigcup\{\neg I_1 \rightarrow I_2\}$; if dwconf$(\neg I_2 \rightarrow I_1) \geqslant$ mc then dwNAR $\leftarrow$ dwNAR$\bigcup\{\neg I_2 \rightarrow I_1\};\}$

end;

end.

Step 9: output dwPAR and dwNAR.

WPNARM-SCCI算法可划分为两个阶段: 第1阶段是Step 1～Step 6, 主要负责从加权数据库WD中挖掘有趣的加权频繁项集和负项集模式; 第2阶段是Step 7～Step 9, 主要负责从频繁项集和负项集中挖掘有效的加权正负关联规则模式, 最后输出规则模式.

算法中子程序功能如下: scan C_1 (WD) 的功能是扫描加权数据库WD, 求出候选1_项集, 计算1_项集的支持数及其权值; scan C_i (WD) 是遍历加权数据库WD, 计算候选i_项集的支持数及其权值.

算法时间复杂性分析如下.

Step 1为初始化, 其时间复杂度可以忽略不计. Step 2和Step 3挖掘1_候选项集C_1和频繁1_项集L_1, 时间复杂度为

$$O(|D| \times |c_1| + |c_1|) \approx O(|D| \times |c_1|),$$

其中$|c_1|$为候选1_项集C_1的个数. Step 4挖掘所有加权频繁i_项集L_i和负i_项集N_i, 总的时间复杂度为

$$O\Big(\sum_{i=2}^{k} |L_{i-1}|^2 \times \sum_{i=2}^{k} |\,c\,|\Big).$$

其中: $|L_{i-1}|$为频繁$(i-1)$_项集L_{i-1}的个数; $|c_i|$为候选i_项集C_i的个数; k为项集长度, 即项集维数. Step 5和Step 6对频繁项集和负项集剪枝, 其时间复杂度为

$$O\Big(\sum_{i=2}^{k} |L_i| + \sum_{i=2}^{k} |N_i|\Big).$$

其中: $|L_i|$为频繁i_项集L_i的个数, $|N_i|$为负i_项集N_i的个数. Step 7挖掘频繁项集中加权强正负关联规则, 其时间复杂性为

$$O\Big(\sum_{i=2}^{k} (|L_i| \times |L_{i(\mathrm{sub})}|)\Big),$$

其中$|L_{i(\mathrm{sub})}|$为频繁i_项集L_i产生的所有子项集(例如$(i-1)$_子项集、$(i-2)$_子项集等)的个数. Step 8挖掘负项集中加权强负关联规则, 其时间复杂性为

$$O\Big(\sum_{i=2}^{k} |N_1| \times |N_{i(\mathrm{sub})}|\Big),$$

其中$|N_{i(\mathrm{sub})}|$为负i_项集N_i产生的所有子项集(例如$(i-1)$_子项集、$(i-2)$_子项集等)的个数. Step 9输出dwPAR和dwNAR中的正负关联规则模式, 其时间复杂度可以忽略不计.

综上所述, WPNARM-SCCI算法的时间复杂性为

$$O\Big(|D|\sum_{i=1}^{k}|c_i| + \sum_{i=2}^{k}|L_{i-1}|^2 + \sum_{i=2}^{k}|L_i| + \sum_{i=2}^{k}|N_i| + \sum_{i=2}^{k}(|L_i| \times |L_{i(\mathrm{sub})}|) + \sum_{i=2}^{k}(|N_i| \times |N_{i(\mathrm{sub})}|)\Big).$$

3　实验与分析

3.1　实验数据及其预处理

实验的硬件环境是: Intel(R) Core(TM) i7-3770 CPU@3.4 GHz 3.4 GHz台式电脑, 内存8.0 G, 硬盘1 T. 软件环境: windows 7+delphli2006+SQL Server 2008.

当前, 无加权正负关联规则挖掘的实验数据集普遍采用关联规则挖掘领域权威的IBM实验室[1]提供的数据生成器生成的人工随机实验数据集, 其命名规则为DxTxRxIx, 或者来自一些网站(例如http://www.kdnuggets.com)的超市数据[7]等. 文本数据是典型的具有项权值变化特征的数据, 很适合作为项权值变化的正负关联模式挖掘实验测试集. 在真实的文本信息数据中进行基于项权值变化的关联模式挖掘实验应该比人工生成的实验数据具有更高的实验研究价值, 因此, 本文选择北京大学网络实验室提供的中文标准数据集CWT200g (http://www.cwirf.org/)的中文文档语料(提取12 024篇)和当前国际上著名文本信息检索标准数据集——日本国家科学信息系统中心信息检索系统测试集NTCIR-5 CLIR(http://research.nii.ac.jp/ntcir/permission/ntcir-5/perm-en-CLIR.html)的Korea_Times2001英文文档语料(该语料有14 069篇英文文档)作为本文实验数据.

实验中, 将文档作为事务, 文档中的特征词作为项目. 中文文档经过中文分词、去除中文停用词和提取中文特征词等预处理. 中文分词程序采用中国科学院计算技术研究所研制编写的汉语词法分析系统ICTCLAS. 英文文档经去除英文停用词和词干提取等

预处理获得英文特征词, 其词干提取程序采用Porter程序(http://tartarus.org/∼martin/PorterStemmer). 特征词权值采用当前文本信息处理中最常用的TF-IDF权值计算方法[22]. 经过文本预处理后, 构建基于向量空间模型的文本数据库和特征词项目库. 文本数据库结构是"事务文档号(DocID), 项目(term), 权值(weight), $\cdots$", 特征词项目库结构是"项目编号(ItemID), 项目(term), 文档频度(df), $\cdots$".

3.2 实验结果及其分析

选择典型的无加权正负关联规则挖掘算法[7](记为PNAR-CPIR), 基于多支持度阈值的无加权正负关联规则挖掘算法[9](记为PNAR-IMLMS)和矩阵加权关联规则挖掘算法MWARM[21]作为对比算法. 将本文算法WPNARM-SCCI与对比算法在中英文标准数据集上分别从支持度、置信度、项目数和测试集数据规模分别变化等进行挖掘性能比较和分析, 最后进行正负关联模式实例分析及其在查询扩展中的应用实例分析. 所涉及的实验参数如下: ms, mc, mi, IN(item number)(所挖掘的项目数量)和n(事务记录总数); NTCIR-5①表示实验中的英文测试集数量为4 936篇, 即$n = 4\,936$; NTCIR-5②表示$n = 14\,069$; 下列表格中, 对于关联规则(association rule, AR), AR 1代表关联规则$A \to B$, AR 2代表$A \to \neg B$, AR 3代表$\neg A \to B$, AR 4代表$\neg A \to \neg B$.

3.2.1 支持度阈值变化时挖掘性能比较

支持度阈值变化时, 4种算法在中英文标准测试集中挖掘候选项集(CI)、频繁项集(FI)、负项集(NI)和正负关联规则的数量结果见表3和表4, 其中实验参数如下: 对于CWT200g, mc = 0.01, mi = 0.005, IN = 50, $n = 12\,024$, ms的取值分别为0.03、0.04、0.05、0.06、0.07、0.08和0.09; 对于NTCIR-5①, mc = 0.01, mi = 0.001, IN = 50, ms为0.07、0.09、0.12、0.14、0.16, PNAR-IMLMS算法的参数是, $\mathrm{ms}(4) = \mathrm{ms}$, $\mathrm{ms}(3) = \mathrm{ms} + 0.004$, $\mathrm{ms}(2) = \mathrm{ms} + 0.006$, $\mathrm{ms}(1) = \mathrm{ms} + 0.008$, $\mathrm{ms}(0) = 0.008$, $\beta = 0.002$, 挖掘到4_项集; 对于NTCIR-5②, mc = 0.01, mi = 0.001, IN = 50, ms为0.009、0.008、0.01、0.02、0.03、0.04、0.05, 挖掘到3_项集, PNAR-IMLMS算法的参数是, $\mathrm{ms}(3) = \mathrm{ms}$, $\mathrm{ms}(2) = \mathrm{ms} + 0.002$, $\mathrm{ms}(1) = \mathrm{ms} + 0.004$, $\mathrm{ms}(0) = 0.005$, $\beta = 0.002$.

表3 支持度变化时4种算法挖掘的项集数量

算法	项集	CWT 200 g	NTCIR-5①	NTCIR-5②
PNAR-CPIR	CI	71 653	37 051	145 116
	FI	13 387	2 721	99 292
	NI	52 521	32 534	43 868
PNAR-IMLMS	CI	33 822	31 372	144 704
	FI	9 872	2 295	99 263
	NI	19 904	27 944	43 537
MWARM	CI	4 584	4 557	4 719
	FI	849	4 504	4 610
WPNARM-SCCI	CI	8 211	17 642	36 032
	FI	740	1 453	4 829
	NI	6 018	15 721	30 488

表4 支持度变化时挖掘的正负关联规则数量

算法	规则	CWT 200 g	NTCIR-5①	NTCIR-5②
PNAR-CPIR	AR1	96 217	7 025	478 803
	AR2	10 172	8 356	11 098
	AR3	9 046	3 095	2 238
	AR4	381 112	114 951	218 045
PNAR-IMLMS	AR1	67 466	5 564	477 140
	AR2	11 386	9 790	30 284
	AR3	11 386	9 790	30 284
	AR4	172 146	101 594	702 200
MWARM	AR1	2 528	3 428	16 624
WPNARM-SCCI	AR1	1 758	3 334	16 720
	AR2	244	1 365	0
	AR3	244	1 365	0
	AR4	19 354	64 480	145 916

3.2.2 置信度阈值变化时挖掘性能比较

在不同的置信度阈值下, 4种算法在中英文标准测试集挖掘正负关联规则的数量结果见表5, 其中实验参数如下: 对于CWT200g, ms = 0.05, mi = 0.005, IN = 50, $n = 12\,024$, mc分别为0.03、0.08、0.09、0.4、0.8、0.9、0.94、0.98; 对于NTCIR-5①, ms = 0.08, mi = 0.001, IN = 50, mc为0.08、0.01、0.03、0.07、0.1、0.3、0.9, 挖掘到4_项集, PNAR-IMLMS的参数是$\mathrm{ms}(0) = 0.008$, $\mathrm{ms}(1) = \mathrm{ms} + 0.008$, $\mathrm{ms}(2) = \mathrm{ms} + 0.006$, $\mathrm{ms}(3) = \mathrm{ms} + 0.004$, $\mathrm{ms}(4) = \mathrm{ms}$, $\beta = 0.002$; 对于NTCIR-5②, ms = 0.02, mi = 0.001, IN = 50, mc为0.01、0.1、0.3、0.5、0.7、0.9、0.95, 挖掘到3_项集, PNAR-IMLMS算法的参数是$\mathrm{ms}(3) = \mathrm{ms}$, $\mathrm{ms}(2) = \mathrm{ms} + 0.002$, $\mathrm{ms}(1) = \mathrm{ms} + 0.004$, $\mathrm{ms}(0) = 0.005$, $\beta = 0.002$.

表5 置信度变化时挖掘的正负关联规则数量

算法	规则	CWT 200 g	NTCIR-5①	NTCIR-5②
PNAR-CPIR	AR1	39 278	12 321	130 721
	AR2	6 259	14 061	4 404
	AR3	778	2 621	416
	AR4	147 641	183 699	20 800
PNAR-IMLMS	AR1	27 083	10 906	204 016
	AR2	11 713	23 437	29 333
	AR3	5 822	19 577	11 321
	AR4	115 175	252 123	553 579
MWARM	AR1	2 310	7 329	685
WPNARM-SCCI	AR1	496	7 191	814
	AR2	366	1 797	0
	AR3	112	1 493	0
	AR4	12 507	138 252	28 055

3.2.3　统计显著性检验分析

采用配对样本t-检验对实验结果进行统计显著性检验分析. 在自由度为7的情况下, 当支持度阈值变化(实验参数同3.2.1节)和置信度阈值分别变化(实验参数同3.2.2节)时, 本文算法与对照算法的候选项集、频繁项集、负项集和正负关联规则数量之间的t-检验值见表6～表8. 其中: 对于表6和表7, $t_{0.05}(6) = 2.447$, $t_{0.01}(6) = 3.707$; 对于表8, $t_{0.05}(7) = 2.306$, $t_{0.01}(7) = 3.355$.

表 6　支持度变化时挖掘的各类项集数量t-检验结果

对比算法	项集	CWT200g		NTCIR-5②	
		t-检验值	结果	t-检验值	结果
WPNARM-SCCI Vs.PNAR-CPIR	CI	3.032	显著	7.226	极显著
	FI	2.429	不显著	5.445	极显著
	NI	3.215	显著	5.912	极显著
WPNARM-SCCI Vs.PNAR-IMLMS	CI	2.989	显著	7.264	极显著
	FI	2.375	不显著	5.438	极显著
	NI	3.645	显著	5.979	极显著
WPNARM-SCCI Vs.MWARM	CI	3.992	极显著	2.338	不显著
	FI	5.272	极显著	2.058	不显著

表 7　支持度变化时挖掘的正负规则数量t-检验值

对比算法	关联规则	CWT200g		NTCIR-5②	
		t-检验值	结果	t-检验值	结果
WPNARM-SCCI Vs.PNAR-CPIR	AR1	2.036	不显著	4.941	极显著
	AR2	4.975	极显著		
	AR3	5.649	极显著		
	AR4	2.536	显著	5.928	极显著
WPNARM-SCCI Vs.PNAR-IMLMS	AR1	1.893	不显著	4.824	极显著
	AR2	3.845	极显著		
	AR3	3.845	极显著		
	AR4	2.200	不显著	8.725	极显著
WPNARM-SCCI Vs.MWARM	AR1	3.504	显著	2.495	显著

表 8　置信度变化时挖掘的正负规则数量t-检验值

对比算法	关联规则	CWT200g		NTCIR-5②	
		t-检验值	结果	t-检验值	结果
WPNARM-SCCI Vs.PNAR-CPIR	AR1	2.899	极显著	1.512	不显著
	AR2	2.916	显著		
	AR3	1.193	不显著		
	AR4	2.177	不显著	3.556	显著
WPNARM-SCCI Vs.PNAR-IMLMS	AR1	3.156	显著	2.085	不显著
	AR2	11.880	极显著		
	AR3	2.491	显著		
	AR4	4.759	不显著	5.486	极显著
WPNARM-SCCI Vs.MWARM	AR1	7.209	极显著	2.695	显著

3.2.4　挖掘时间效率比较

在支持度阈值变化(实验参数同3.2.1节)和置信度阈值变化(实验参数同3.2.2节)两种情况下, 本文算法WPNARM-SCCI和对比算法PNAR-CPIR、PNAR-IMLMS在中英文测试集中挖掘项集和关联规则时间如表9所示.

表 9　3种算法挖掘项集和关联规则的时间　　s

算法	不同支持度下挖掘项集和关联规则时间		不同置信度下挖掘关联规则时间	
	CWT200g	NTCIR-5②	CWT200g	NTCIR-5②
PNAR-CPIR	27 456	37 727	15 684	11 457
PNAR-IMLMS	9 813	47 303	5 292	25 133
WPNARM-SCCI	1 295	16 864	244	1 342

3.2.5　算法的扩展性能分析

为了测试算法的可扩展性能, 在项目数量和文档规模分别变化情况下, 本文算法WPNARM-SCCI在测试集CWT200g中挖掘项集和正负关联规则数量结果如图1～图4所示. 其中实验参数为ms = 0.05, mc = 0.07, mi = 0.001.

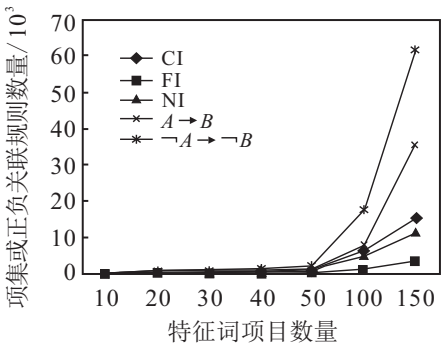

图 1　不同项目数的项集和正负规则数量变化

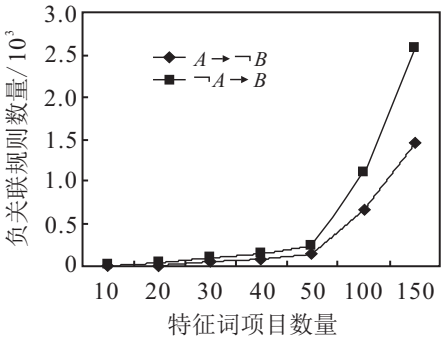

图 2　不同项目数的负关联规则数量变化

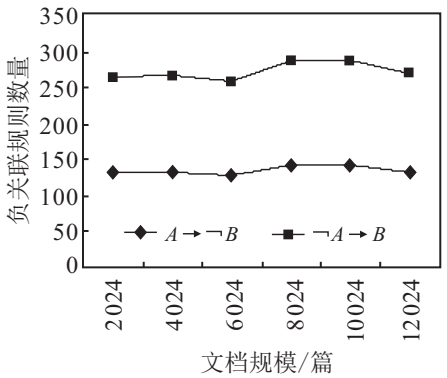

图 3　不同文档规模的负关联规则数量变化

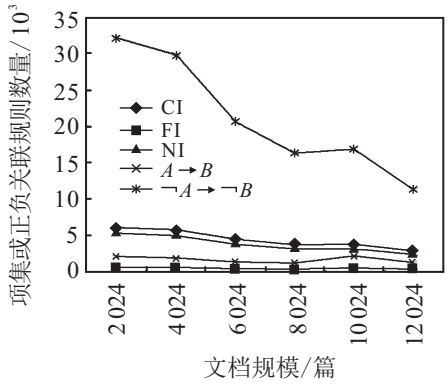

图 4 不同文档规模的项集和正负规则数量变化

3.2.6 项集的剪枝性能

在兴趣度阈值变化和支持度阈值变化两种情况下, 本文所提出的WPNARM-SCCI算法在中文测试集CWT200g挖掘频繁项集(FI)和负项集(NI)的数量如表10和表11所示. 表10中$\mathrm{mi}=0$表示不剪枝的挖掘结果; 表11中MP表示采用剪枝策略(MP), MNP表示不进行项集剪枝(MNP).

表 10 兴趣度变化情况下项集的剪枝结果
$(\mathrm{ms}=0.05, \mathrm{IN}=100, n=12\,024)$

mi	FI	降幅/%	NI	降幅/%
0	985	0	5 139	0
0.000 1	979	−0.61	5 035	−2.02
0.003 0	979	−0.61	4 339	−15.57
0.005 0	979	−0.61	3 850	−25.08
0.030 0	979	−0.61	623	−87.88
0.050 0	750	−23.86	0	0
0.070 0	275	−72.08	0	0
0.100 0	100	−89.85	0	0

表 11 支持度变化情况下项集的剪枝结果
$(\mathrm{mi}=0.01, \mathrm{IN}=50, n=12\,024)$

ms	MNP		MP	
	FI	NI	FI	NI
0.03	5 008	15 568	4 968	12 535
0.04	1 807	6 752	1 793	4 649
0.05	985	5 139	979	3 049
0.06	622	4 808	620	2 781
0.07	264	3 218	262	2 000
0.08	156	1 558	155	1 007
0.09	17	219	17	192
合计	8 859	37 262	8 794	26 213

3.2.7 正负关联模式实例分析

取20个特征词项目, 即“部分, 部门, 采用, 参加, 参与, 产品, 产生, 长期, 超过, 成本, 成长, 成功, 城市, 程度, 出现, 处理, 传奇, 传统, 创造, 存在”, 本文所提出的WPNARM-SCCI算法和对比算法PNAR-CPIR、PNAR-IMLMS在中文文本数据集CWT200g上进行挖掘. 在挖掘结果中, 提取一些典型的、与特征词“参与”相关的频繁2_项集、负2_项集和部分以“参与”为前件的正负关联规则实例进行分析, 实例结果如表12所示. 其中: PAR表示正关联规则(PAR), NAR表示负关联规则(NAR).

表12的实例结果表明, 含“参与”的频繁2_项集和负2_项集以及以“参与”为前件正负关联规则实例中, 本文算法WPNARM-SCCI挖掘的频繁项集和负项集, 以及正负关联规则模式比对比算法挖掘的更接近实际情况, 避免了无效的和虚假的关联模式产生. 例如, 特征词“参与”和“参加”是近义词, 在一句话或者一段话中应该很少同时出现, 所以, 项集“{参与, 参加}”应该是负项集, 关联规则“参与→参加”不应该是强关联规则, 而应该是强负关联规则“参与→¬参加”或“¬参与→参加”. 实验结果表明: 本文算法WPNARM-SCCI能挖掘出感兴趣的负项集“{参与, 参加}”和强负关联规则“参与→¬参加”或“¬参与→参加”, 没有挖掘出形如“参与→参加”等这类无效和虚假的模式; 而对比算法挖掘出的项集“{参与, 参加}”是频繁项集, 并挖掘出强正关联规则“参与→参加”, 而这类关联模式应该是虚假的、无趣的和无效的模式.

负项集和负关联规则表明是一种具有否定关联的关系, 对于负项集“{参加, 参与}”, 其所表示的实际意义是, “参加”、“参与”这2个特征词不会同时出现的可能性很高, 对于负关联规则{参与→¬参加}所表示的意义是: 当特征词“参与”出现时, 特征词“参加”不会出现的可能性很大. 频繁项集和正关联规则表明是一种正相关的关联, 当“{参加, 参与}”是频繁项集时, 所表示的含义是“参加”、“参与”这2个特征词同时出现的可能性很高, 对于其正关联规则{参与→参加}所表示的意义是, 当特征词“参与”出现时, 特征词“参加”同时出现的可能性很大.

3.2.8 正负关联规则应用实例分析

正负关联规则模式在信息检索查询扩展、文本分类和聚类等领域具有很高的应用价值和实际意义. 例如, 通过文本特征词关联模式挖掘得到的正关联规则模式为信息检索查询扩展提供丰富和可靠的扩展词来源, 然而, 所得到的扩展词中有些可能是一些虚假的和无效的, 通过负关联规则模式挖掘可以发现那些无效和虚假的扩展词. 对此, 本文进行基于权值变化的正负关联模式挖掘应用实验研究, 以当前著名的信息检索系统标准测试集NTCIR-5 CLIR_Korea_Times 2001为实验数据测进行试集, 该数据集包含标准测试文档集、查询集和结果集.

表 12　含“参与”的2_项集和以“参与”为前件正负关联规则实例
(CWT200g : ms = 0.02, mc = 0.01, mi = 0.005, n = 12 024)

算法	项集	项集和正负关联规则实例
PNAR-CPIR	FI	{参与, 产品}, {参与, 产生}, {参与, 长期}, {参与, 成本}, {参与, 成功}, {参与, 程度}, {参与, 出现}, {参与, 处理}, {参与, 传统}, {参与, 创造}, {参与, 存在}, {部分, 参与}, {**参加**, **参与**}, {采用, 参与}, {参与, 超过}
	NI	{参与, 传奇}
	PAR	{参与→采用}, {参与→产品}, {参与→超过}, {参与→成本}, {参与→程度}, {参与→出现}, {参与→创造}, {参与→存在}, {参与→部分}, {参与→部门}, {**参与→参加**}, {参与→产生}, {参与→长期}, {参与→成功}, {参与→处理}, {参与→传统}
	NAR	{参与→¬传奇}, {¬参与→传奇}
PNAR-IMLMS	FI	{参与, 产品}, {参与, 产生}, {参与, 长期}, {参与, 成本}, {参与, 成功}, {参与, 程度}, {参与, 出现}, {参与, 处理}, {参与, 传统}, {参与, 创造}, {参与, 存在}, {部分, 参与}, {部门, 参与}, {参与, 采用}, {**参加**, **参与**}, {部门, 参与}, {参与, 超过}
	NI	{参与, 传奇}
	PAR	{参与→采用}, {参与→部门}, {参与→存在}, {参与→创造}, {参与→传统}, {参与→处理}, {参与→出现}, {参与→程度}, {参与→成功}, {参与→成本}, {参与→长期}, {参与→产生}, {参与→产品}, {参与→部分}, {参与→超过}, {**参与→参加**}
	NAR	{¬参与→¬部分}, {¬参与→¬部门}, {¬参与→¬采用}, {¬参与→¬产生}, {¬参与→¬产品}, {¬参与→¬参加}, {¬参与→¬存在}, {¬参与→¬创造}, {¬参与→¬传统}, {¬参与→¬处理}, {¬参与→¬出现}, {¬参与→¬程度}, {¬参与→¬成功}, {¬参与→¬成本}, {¬参与→¬超过}, {¬参与→¬长期}, {参与→¬传奇}, {¬参与→传奇}
WPNARM-SCCI	FI	{参与, 产品}, {参与, 产生}, {参与, 长期}, {参与, 成本}, {参与, 成功}, {参与, 程度}, {参与, 出现}, {参与, 处理}, {参与, 传统}, {参与, 创造}, {参与, 存在}, {部分, 参与}, {部门, 参与}
	NI	{**参与**, **参加**}, {参与, 超过}
	PAR	{参与→采用}, {参与→部门}, {参与→存在}, {参与→创造}, {参与→传统}, {参与→处理}, {参与→出现}, {参与→程度}, {参与→成功}, {参与→成本}, {参与→长期}, {参与→产生}, {参与→产品}, {参与→部分}
	NAR	{¬参与→¬部门}, {**参与→¬参加**}, {**¬参与→参加**}, {¬参与→¬部分}, {¬参与→¬存在}, {¬参与→¬创造}, {¬参与→¬传统}, {¬参与→¬处理}, {¬参与→¬出现}, {¬参与→¬程度}, {¬参与→¬成功}, {¬参与→¬成本}, {¬参与→¬长期}, {¬参与→¬产生}, {¬参与→¬产品}, {¬参与→¬城市}, {¬参与→¬成长}, {¬参与→¬采用}, {¬参与→超过}, {¬参与→成长}

实验方法是: 首先从NTCIR-5 CLIR标准查询集中选择5个英文查询对Korea_Times 2001英文测试文档集进行向量空间模型检索, 根据查询和文档相似度阈值提取初检文档; 然后, 采用本文挖掘算法WPNARM-SCCI对初检文档挖掘与原查询相关的频繁项集、负项集和正负特征词关联规则, 并从正关联规则中提取正扩展词, 从负关联规则中提取负扩展词; 最后, 将正扩展词中含有的负扩展词删除后, 得到最终的扩展词与原查询组合成新查询进行第2次检索. 实验中的5个英文原查询的编号是001, 003, 025, 040, 047, 查询内容为其title部分. 以查全率和查准率为检索评价指标, 分初检 (记为baseline)、加入正扩展词后的检索 (记为P_EX)、正扩展词中减去负扩展词后的检索 (P-N_EX) 等3种情况进行实验, 以NTCIR结果集中的“Relaxed”相关性统计其查全率 (recall, 记为Rec) 和查准率 (precision, 记为Pre), 初检相似度阈值设为0.1, 挖掘参数为ms = 0.05, mc = 0.01, mi = 0.001, 通过本文挖掘算法获得的最终扩展词 (即正扩展词中除去负扩展词后得到的扩展词) 数量以及检索实验结果如表13～表15所示.

表 13　各查询的扩展词数量

查询号	正扩展词	负扩展词	最终扩展词
001	39	10	29
003	71	11	60
025	39	25	14
040	59	2	57
047	30	4	28

表 14　各查询的查全率和查准率

查询号	查询与文档相似度	baseline		P_EX		P-N_EX	
		Rec	Pre	Rec	Pre	Rec	Pre
001	0.3	0.40	0.7	0.6	0.53	0.7	0.6
	0.4	0	0	0.4	1.0	0.6	1.0
003	0.3	0	0	0.45	0.59	0.5	0.57
	0.4	0	0	0.1	0.25	0.15	0.25
025	0.3	0.4	0.6	0.53	0.67	0.64	0.7
	0.4	0.33	0.67	0.43	1.0	0.53	0.67
040	0.3	0.44	1.0	0.56	1.0	0.56	1.0
	0.4	0.22	1.0	0.33	1.0	0.33	1.0
047	0.3	0	0	0.55	0.46	0.55	0.44
	0.4	0	0	0.2	0.56	0.21	0.57

表 15 部分正负关联规则和扩展词实例(查询号 001)

关联模式	数量	部分实例
正关联规则 $(A \to B)$	55	warner $\to$ aol, onlin $\to$ industry, onlin $\to$ content, onlin $\to$ retail, merger $\to$ expect, merger $\to$ telecom, merger $\to$ oper, merger $\to$ synergy, merger $\to$ trillion, $\cdots$
负关联规则 $(A \to \neg B)$	27	american $\to \neg$ trillion, merger $\to \neg$ retail, warner $\to \neg$ content, $\cdots$
正扩展词	39	aol, industry, content, retail, expect, telecom, oper, synergy, trillion, $\cdots$
负扩展词	10	trillion, retail, content, $\cdots$
最终扩展词	29	aol, industry, expect, telecom, oper, synergy, $\cdots$

表 14 的实验结果表明, 通过本文挖掘算法获得的与原查询相关的扩展词可以改善信息检索性能, 当去除负扩展词后, 检索性能改善得更好.

以原查询号“001”为例, 在挖掘结果中列举以该查询词为前件的正负关联规则模式及其正负扩展词和最终扩展词结果的部分实例(限于篇幅没有全部列出)如表 15 所示, 这些模式都是基于项权值变化挖掘的模式, 001 号原查询 title 内容是“time warner, American online(AOL), merger, impact”.

3.2.9 实验结果分析

上述实验结果表明, 与 3 种典型的无加权和加权对比算法相比较, 本文 WPNARM-SCCI 算法具有如下特点:

1) 表 3 ~ 表 5 的结果表明, 本文 WPNARM-SCCI 算法挖掘的项集和正负规则模式数量都比无加权挖掘算法 PNAR-IMLMS、PNAR-CPIR 挖掘的少, 降幅比较大. 与现有基于权值变化的加权挖掘算法 MWARM 相比较, 本文算法挖掘的频繁项集与正关联规则数量相当, 而挖掘的候选项集数量增多, 主要原因是 MWARM 算法主要对候选项集剪枝, 而本文算法是对频繁项集和负项集剪枝. 从表 6 ~ 表 8 可以看出, 在自由度为 7 的情况下, 本文算法实验结果与对照算法的结果之间存在显著性差异. 其中: 对于无加权对照算法, 其候选项集、负项集以及负关联规则 $A \to \neg B$ 和 $\neg A \to B$ 的显著性水平达到“显著”或者“极显著”; 对于加权对照算法, 其候选项集、频繁项集和正关联规则的显著性水平达到“显著”或者“极显著”, 这些结果表明本文算法挖掘性能的提高在统计上有一定意义.

2) 表 9 ~ 表 11 表明, 本文算法挖掘时间最少, 挖掘效率得到提高, 采用 SCCI 框架, 项权值变化的挖掘获得了良好的剪枝性能.

3) 由图 1 ~ 图 4 的结果可知, 随着项目数的增多和文档规模的增大, 本文算法表现出良好的可扩展性.

4) 实例分析表明, 项权值变化的正负关联规则模式挖掘对比算法能挖掘到有趣的关联模式, 避免一些无趣和无效模式的出现, 能改善信息检索查询性能, 在信息检索查询扩展中具有较高的应用价值和实际意义.

主要原因分析如下: 本文 WPNARM-SCCI 算法属于基于项权值变化的加权关联规则挖掘算法, 考虑了项集权值依赖于事务并随事务记录变化而变化的加权数据特点, 其挖掘出来的关联模式更加合理, 更接近实际; 采用新的剪枝策略, 减少了大量的无趣的和冗余的关联模式产生, 使得关联模式数量大大减少, 极大地提高了挖掘效率. 对比算法 PNAR-CPIR、PNAR-IMLMS 属于基于频度的项无加权挖掘算法, 没有考虑项集权重, 产生很多无趣的和虚假的项集和关联规则, 使得频繁项集和关联规则数量增多, 导致挖掘时间增加, 其挖掘效率大大减低. 对比算法 MWARM 虽然是基于权值变化的挖掘算法, 但不能挖掘负项集和负关联规则.

实验结果还表明, 本文算法在不同的数据集中, 挖掘结果还存在差异, 有些结果在统计上没有显著性差异, 表现出挖掘性能还不稳定, 值得进一步研究和探讨. 例如, 在英文文档数为 4 936 时可以挖掘出正负关联规则, 但文档数为 14 069 时, 负规则 $\neg A \to B$ 和 $A \to \neg B$ 就没能挖掘出来. 上述问题的主要原因是数据集的权值大小和分布不同, 当项集的总体权值偏小时, 有些负规则就挖掘不出来.

4 结 论

基于项权值变化的挖掘研究在文本挖掘、信息检索等领域具有重要的理论和应用价值, 特别是其加权负关联模式的作用和应用价值日益凸显. 本文深入研究了项权值变化的数据模型, 给出其数据模型的形式化表示, 提出了新的加权关联模式剪枝策略, 构建了新的加权关联模式评价框架 SCCI, 提出了基于项目权值变化和 SCCI 评价框架的加权正负关联规则挖掘算法 WPNARM-SCCI, 有效地解决了基于项权值变化的加权负模式挖掘问题, 并通过实验结果表明了该算法的有效性. 下一步研究的重点是, 研究项目权值大小与分布对挖掘关联模式的影响规律, 努力探索将该成果运用于信息检索、跨语言信息检索等领域, 提高信息检索查询性能, 以及将其运用于教育数据挖掘, 以发现更多、更合理的教育教学模式.

参考文献(References)

[1] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]. Proc of the 20th Int Conf on Very

Large Data Bases(VLDB). Santiago, 1994: 487-499.

[2] Narmadha D, NaveenSundar G, Geetha S. A novel approach to prune mined association rules in large databases[C]. Proc of the 3rd IEEE Int Conf on Electronics Computer Technology. Kanyakumari, India: IEEE, 2011: 409-413.

[3] 董杰, 韩敏. 挖掘事务间频繁闭项集的高效率算法[J]. 控制与决策, 2008, 23(9): 994-998.
(Dong J, Han M. Efficient algorithm of mining frequent closed inter-transaction itemsets[J]. Control and Decision, 2008, 23(9): 994-998.)

[4] Glass D H. Confirmation measures of association rule interestingness[J]. Knowledge-Based Systems, 2013, 44(5): 65-77.

[5] 何波. 基于频繁模式树的分布式关联规则挖掘算法[J]. 控制与决策, 2012, 27(4): 619-622.
(He B. Distributed algorithm for mining association rules based on FP-tree[J]. Control and Decision, 2012, 27(4): 619-622.)

[6] Shaheen M, Shahbaz M, Guergachi A. Context based positive and negative spatio-temporal association rule mining[J]. Knowledge-Based Systems, 2013, 37(1): 261-273.

[7] Wu X, Zhang C Q, Zhang S C. Efficient mining of both positive and negative association rules[J]. ACM Trans on Information Systems, 2004, 22(3): 381-405.

[8] Taniar D, Rahayu W, Daly O, et al. Mining hierarchical negative association rules[J]. Int J of Computational Intelligence Systems, 2012, 5(3): 434-451.

[9] Swesi I M A O, Bakar A A, Kadir A S A. Mining positive and negative association rules from interesting frequent and infrequent itemsets[C]. Proc of the 9th IEEE Int Conf on Fuzzy Systems and Knowledge Discovery(FSKD 2012). Chengdu: IEEE Computer Society, 2012: 650-655.

[10] Hämäläinen W. Kingfisher: An efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures[J]. Knowledge and Information Systems, 2012, 32(2): 383-414.

[11] Bhargava R, Lade S. Effective positive negative association rule mining using improved frequent pattern tree[J]. Int J of Advanced Research in Computer Science and Software Engineering, 2013, 3(4): 193-199.

[12] Cai C H, da A, Fu W C, et al. Mining association rules with weighted items[C]. Proc of Int Database Engineering and Application Symposiums. Cardiff: IEEE, 1998: 68-77.

[13] Yun U, Ryu K H. Approximate weighted frequent pattern mining with/without noisy environments[J]. Knowledge-Based Systems, 2011, 24 (1): 73-82.

[14] Pears R, Koh Y S. Weighted association rule mining using particle swarm optimization[C]. Proc of PAKDD 2011 Workshops. Berlin Heidelberg: Springer-Verlag, 2012: 327-338.

[15] Tan J. Weighted association rules mining algorithm research[J]. Applied Mechanics and Materials, 2013(241/242/243/244): 1598-1601.

[16] Vo B, Coenen F, Le B. A new method for mining frequent weighted itemsets based on WIT-trees[J]. Expert Systems with Applications, 2013, 40(4): 1256-1264.

[17] Pears R, Koh Y S, Dobbie G, et al. Weighted association rule mining via a graph based connectivity model[J]. Information Sciences, 2013, 218(1): 61-84.

[18] Jiang H, Luan X, Dong X J. Mining weighted negative association rules from infrequent itemsets based on multiple supports[C]. Proc of the 2012 Int Conf on Industrial Control and Electronics Engineering. Xi'an: IEEE Computer Society, 2012: 89-92.

[19] Zhao Y Y, Jiang H, Geng R, et al. Mining weighted negative association rules based on correlation from infrequent items[C]. Proc of the 2009 Int Conf on Advanced Computer Control. Singapore: IEEE Computer Society, 2009: 270-273.

[20] 谭义红, 林亚平. 向量空间模型中完全加权关联规则的挖掘[J]. 计算机工程与应用, 2003, 39(13): 208-211.
(Tan Y H, Lin Y P. Mining all-weighted association rules from vector space model[J]. Computer Engineering and Applications, 2003, 39(13): 208-211.)

[21] 黄名选, 严小卫, 张师超. 基于矩阵加权关联规则挖掘的伪相关反馈查询扩展[J]. 软件学报, 2009, 20(7): 1854-1865.
(Huang M X, Yan X W, Zhang S C. Query expansion of pseudo relevance feedback based on matrix-weighted association rules mining[J]. J of Software, 2009, 20(7): 1854-1865.)

[22] Stefan Büttcher, Charles L A Clarke, Gordon V Cormack. Information retrieval implementing and evaluating search engines[M]. Beijing: China Machine Press, 2012: 40-41.

[23] 沈斌, 姚敏. 关联且项项正相关频繁模式挖掘[J]. 浙江大学学报: 工学版, 2009, 43(12): 2171-2185.
(Shen B, Yao M. Mining associated and item-item correlated frequent patterns[J]. J of Zhejiang University: Engineering Science, 2009, 43(12): 2171-2185.)

[24] 周皓峰, 朱扬勇, 施伯乐. 一个基于兴趣度的关联规则采掘算法[J]. 计算机研究与发展, 2002, 39(4): 627-633.
(Zhou H F, Zhu Y Y, Shi B L. Amining algorithm for association rules based on interest measure[J]. J of Computer Research and Development, 2002, 39(4): 627-633.)

(责任编辑：曹洪武)