基于项权值变化的完全加权正负关联规则挖掘

周秀梅[1],黄名选[2]
(1.南宁地区教育学院数学与计算机科学系,广西崇左 532200; 2.广西财经学院信息与统计学院,广西南宁 530003)

摘 要: 本文提出一种基于项权值变化的完全加权正负关联规则挖掘算法,解决了基于项权值变化的负模式挖掘问题.该算法考虑项权值依赖于事务记录的特点,采用新的项集剪枝方法和模式评价框架,通过项集的项内权值比和维数比的简单计算和比较,挖掘有效的完全加权正负关联规则.实验结果表明,与现有无加权正负关联规则挖掘算法比较,本文算法能避免无效的模式出现,其挖掘时间和候选项集数量明显减少,减幅最大分别可达 94.09% 和 88.16%.

关键词: 数据挖掘; 完全加权关联规则; 负关联规则; 频繁项集
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2015)08-1545-10
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.08.012

All-Weighted Positive and Negative Association Rules Mining Based on Dynamic Item Weight

ZHOU Xiu-mei[1], HUANG Ming-xuan[2]
(1. *Department of Mathematics and Computer Science, Nanning Prefecture Education College, Chongzuo, Guangxi* 532200, *China*;
2. *School of Information and Statistics, Guangxi University of Finance and Economics, Nanning, Guangxi* 530003, *China*)

Abstract: This paper proposes a mining algorithm of all-weighted positive and negative association rules based on dynamic item weight, which can solve the problems of negative patterns mining based on dynamic item weight. This algorithm took the dynamic item weight dependent on transaction records into consideration, and adopted the itemset pruning method and pattern evaluation framework so as to discover effective all-weighted positive & negative association rules via simple calculation and comparison of weight ratio and dimension ratio from the itemset. The experimental results show that this algorithm can prevent ineffective patterns, which makes the maximal declines of the mining time and number of the candidate itemsets by up to 94.09% and 88.16% respectively compared with the existing unweighted positive and negative association rule mining algorithms.

Key words: data mining; all-weighted association rule; negative association rule; frequent itemset

1 引言

关联规则挖掘是数据挖掘中一个重要的研究内容.近 20 年来,关联规则挖掘研究主要集中在基于项目频度的挖掘和基于项权值的挖掘两个方面.早期的关联规则挖掘技术是基于项频度的挖掘,其特点是以项目在数据库中出现的频度作为挖掘关联模式的依据,按平等一致的方式处理项目,其典型算法是 1993 年 AGRAWAL 等提出的 Apriori 算法[1].此后,众多学者从不同角度提出了改进算法,有的从剪枝策略上进行改进,例如,宋威等采用新剪枝策略,提出一种新的频繁项集挖掘算法[2],Narmadha 等提出一种新的关联规则剪枝策略[3],有效地提高关联规则挖掘效率;有的在关联规则的评价方式上进行改进,GLass 提出 2 种新的关联规则兴趣度确定方法[4],解决项集生成的瓶颈问题;有的从改进挖掘方式进行研究,耿生玲等将包含度引入软集数据关联规则挖掘,提出了基于软集包含度理论的关联规则挖掘算法[5],Shaheen 等提出基于上下文的时空关联规则挖掘算法[6],均获得良好的挖掘效果.1997 年以来,基于项频度的负关联规则挖掘得到了讨论和研究,Wu 等提出一种新的正负关联规则挖掘算法[7],以及 SWESI 等提出的基于多支持度的正负关联规则挖掘算法[8],BHARGAVA

收稿日期:2014-03-20;修回日期:2014-08-18;责任编辑:覃怀银
基金项目:国家自然科学基金(No.61262028,No.61363037);广西自然科学基金(No.2012GXNSFAA053235);广西财经学院数量经济学创新团队项目(No.2014CX01)

等提出的基于频繁模式树的正负关联规则挖掘算法[9],均是基于项频度的正负关联规则挖掘的典型算法.基于项频度挖掘的缺陷是:只考虑项频度,忽略项权值,导致大量无效的关联模式产生.

基于项权值的模式挖掘能有效地克服基于频度挖掘的缺陷,其特点是引入项目权值,以体现项目之间具有不同的重要性,其研究包括基于项权值固定的正负模式挖掘和项权值变化的关联模式挖掘.基于项权值固定的正关联规则挖掘研究起于 1998 年 Cai 等提出的加权关联规则挖掘算法[10].在此基础上,Yun 等提出在噪音环境下稳定地挖掘加权频繁项集的挖掘算法[11],Pears 等使用粒子群优化技术分配项集的权值,提出基于粒子群优化的加权关联规则挖掘算法[12],以及 TAN 等提出的加权关联规则算法[13]等都取得了良好的挖掘性能.随着研究的深入,加权负关联规则的作用日显突出,在挖掘有利因素的同时也期望发现一些不利因素,通过负关联规则分析可以达到此目的,2008 年以来,基于项权值固定的加权负关联模式挖掘得到重视和研究,文献[14]提出从非频繁项集挖掘加权负关联规则算法,获得良好的挖掘效果.

基于项权值固定的挖掘存在的缺陷是:忽略项目在事务记录中具有不同权值的情况.基于项权值变化的挖掘有效地解决了该缺陷,其特点是不仅引入项目权值,还考虑项目在不同的事务记录中具有不同的权值.将项目权值依赖于事务并随事务记录变化的数据(即项权值变化的数据)称为完全加权数据,也称矩阵加权数据.基于项权值变化的完全加权关联规则挖掘研究起于 2003 年谭义红等提出的 KWEstimate 算法[15].此后,黄名选等提出完全加权词间关联规则挖掘算法[16]和面向查询扩展的矩阵加权关联规则挖掘算法[17],刘远超等提出基于粗集理论的中文关键词短语构成规则挖掘算法[18],这些算法有效地克服基于项权值固定的关联模式挖掘缺陷,解决了基于项权值变化的正关联模式挖掘技术问题.

当前,基于项频度的挖掘和基于项权值固定的挖掘得到深入而广泛地研究,基于项权值变化的关联模式挖掘研究并不深入,其国内外相关的文献不多.现有基于项权值变化的挖掘算法[15~18]没能解决基于项权值变化的负关联模式挖掘问题.鉴于此,本文提出一种基于项权值变化的完全加权正负关联规则挖掘方法.该方法考虑了项权值依赖于事务记录的特点,构建一种完全加权正负关联模式评价框架 SCPIRCI(Support-Conditional Probability Increment Ratio-Correlation-Interest),提出一种完全加权项集剪枝策略,设计基于 SCPIRCI 评价框架的完全加权正负关联规则挖掘算法.实验结果表明,与经典的挖掘算法比较,本文算法挖掘的候选项集数量和挖掘时间明显减少,能避免无效的关联模式产生,挖掘效率得到极大提高,有效地解决了基于项权值变化的负关联规则挖掘问题.

2 基于项权值固定的和项权值变化的数据模型比较

基于项权值固定的数据和基于项权值变化的数据主要区别是:(1)项目权值设置方式和来源不同.前者的项目权值由用户主观设置,独立于事务,在挖掘过程中固定不变;后者的项目权值依赖于事务记录,随事务记录不同而变化,其权值的计算是根据具体完全加权数据的项目权值计算方法.例如,文本数据库中特征词项目权值采用 TF-IDF(Term Frequency Imverse Document Frequency)权值计算方法[19,20];(2)数据模型不同,如表 1 和表 2 所示,其中,$\{T_1, T_2, \cdots, T_n\}$是事务集合,$\{i_1, i_2, \cdots, i_m\}$是其项目集合.在基于项权值固定的加权数据模型中,$\{w_1, w_2, \cdots, w_m\}$是相应项目的固定权值,"1/0"的"1"表示项目在事务记录中出现,"0"表示不出现.在基于项权值变化的完全加权数据模型中,"$w[T_i][i_j]/0(1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m)$"为项目在各个事务记录的权值.

表 1 项加权数据模型

事务	$i_1:w_1$	$i_2:w_2$	…	$i_m:w_m$
T_1	1/0	1/0	…	1/0
T_2	1/0	1/0	…	1/0
…	…	…	…	…
T_n	1/0	1/0	…	1/0

表 2 项完全加权数据模型

事务	i_1	i_2	…	i_m
T_1	$w[T_1][i_1]/0$	$w[T_1][i_2]/0$	…	$w[T_1][i_m]/0$
T_2	$w[T_2][i_1]/0$	$w[T_2][i_2]/0$	…	$w[T_2][i_m]/0$
…	…	…	…	…
T_n	$w[T_n][i_1]/0$	$w[T_n][i_2]/0$	…	$w[T_n][i_m]/0$

3 基本概念

设基于项权值变化的完全加权数据库(All-Weighted Database, AWD)AWD = $\{T_1, T_2, \cdots, T_n\}$,事务数为 n,$T_i(1 \leqslant i \leqslant n)$为 AWD 中的第 i 个事务,项集 $I = \{i_1, i_2, \cdots, i_m\}$表示 AWD 中全部项目集合,项目数为 m,$i_j(1 \leqslant j \leqslant m)$为 AWD 中第 j 个项目,$w[T_i][i_j](1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m)$为项目 i_j在事务记录 T_i中的权值,如表 2 所示.设 I_1, I_2是项集 I 的子项集,$I_1 \subset I$,$I_2 \subset I$ 且,$I_1 \cup I_2 = I$,$I_1 \cap I_2 = \varnothing$,给出如下基本定义:

定义 1 对于完全加权项集 I,其完全加权支持度(all-weighted support, awsup)的计算公式[15]如式(1)所示,其中,k 为项集 I 的长度(即 I 的项目个数).

$$\text{awsup}(I)=\frac{\sum_{T_i\in(\text{AWD})}\sum_{i_j\in I}w[T_i][i_j]}{n\times k}=\frac{w_I}{n\times k}\quad(1)$$

完全加权负项集(I_1,I_2)的支持度计算公式如式(2)~(5)所示.

$$\text{awsup}(\neg I)=1-\text{awsup}(I)\quad(2)$$

$$\text{awsup}(I_1,\neg I_2)=\text{awsup}(I_1)-\text{awsup}(I_1,I_2)\quad(3)$$

$$\text{awsup}(\neg I_1,I_2)=\text{awsup}(I_2)-\text{awsup}(I_1,I_2)\quad(4)$$

$$\text{awsup}(\neg I_1,\neg I_2)=1-\text{awsup}(I_1)-\text{awsup}(I_2)+\text{awsup}(I_1,I_2)\quad(5)$$

定义 2　设最小支持度阈值为 ms(minimum support),对于完全加权项集 I,若 $\text{awsup}(I)\geqslant\text{ms}$,则称项集 I 为完全加权频繁项集.对于完全加权项集(I_1,I_2),当 I_1和 I_2都是频繁项集时,若 $\text{awsup}(I_1,I_2)<\text{ms}$,则项集$(I_1,I_2)$称为完全加权负项集.

定义 3　基于无加权数据挖掘环境下的兴趣度模型定义[21],对于完全加权项集(I_1,I_2),给出其完全加权项集兴趣度(all-weighted Itemset Interest,awII)计算公式如式(6)~(9)所示:

$$\text{awII}(I_1,I_2)=\text{awsup}(I_1)\times\text{awsup}(I_1,I_2)\times(1\quad\text{awsup}(I_2))\quad(6)$$

$$\text{awII}(I_1,\neg I_2)=\text{awsup}(I_1)\times\text{awsup}(I_2)\times(\text{awsup}(I_1)-\text{awsup}(I_1,I_2))\quad(7)$$

$$\text{awII}(\neg I_1,I_2)=(1-\text{awsup}(I_1))\times(1-\text{awsup}(I_2)\times(\text{awsup}(I_2)-\text{awsup}(I_1,I_2))\quad(8)$$

$$\text{awII}(\neg I_1,\neg I_2)=\text{awsup}(I_2)\times(1-\text{awsup}(I_1))\times(1-\text{awsup}(I_1)-\text{awsup}(I_2)+\text{awsup}(I_1,I_2))\quad(9)$$

定义 4　CPIR(Conditional-Probability Increment Ratio)模型是用条件概率和先验概率的比值来表达 $p(I_2/I_1)$相对 $p(I_2)$的递增程度[7].将 CPIR 模型思想应用于完全加权数据挖掘,给出完全加权 CPIR(all-weighted CPIR,awCPIR)计算公式如式(10)~(13)所示.本文将 awCPIR 值作为完全加权关联规则的置信度.

$$\text{awCPIR}(I_1\rightarrow I_2)=\frac{\text{awsup}(I_1,I_2)-\text{awsup}(I_1)\text{awsup}(I_2)}{\text{awsup}(I_1)(1-\text{awsup}(I_2))}\quad(10)$$

$$\text{awCPIR}(I_1\rightarrow\neg I_2)=\frac{\text{awsup}(I_1)\text{awsup}(I_2)-\text{awsup}(I_1,I_2)}{\text{awsup}(I_1)\text{awsup}(I_2)}\quad(11)$$

$$\text{awCPIR}(\neg I_1\rightarrow I_2)=\frac{\text{awsup}(I_1)\text{awsup}(I_2)-\text{awsup}(I_1,I_2)}{(1-\text{awsup}(I_1))(1-\text{awsup}(I_2))}\quad(12)$$

$$\text{awCPIR}(\neg I_1\rightarrow\neg I_2)=\frac{\text{awsup}(I_1,I_2)-\text{awsup}(I_1)\text{awsup}(I_2)}{(1-\text{awsup}(I_1))\text{awsup}(I_2)}\quad(13)$$

定义 5　设 w_{12}和 w_1、w_2分别为完全加权项集(I_1,I_2)及其子项集 I_1和 I_2在完全加权数据库 AWD 中的权值总和,将 w_{12}和$(w_1\times w_2)$的比值称为完全加权项集权值比率(all-weighted Itemset Weight Ratio,awIWR),简称项内权值比,即式(14)所示.

$$\text{awIWR}(I_1,I_2)=\frac{w_{12}}{w_1\times w_2}\quad(14)$$

定义 6　设 k_{12},k_1和 k_2分别为项集(I_1,I_2)及其子项集 I_1和 I_2的项目个数,将 k_{12}和$(k_1\times k_2)$的比值称为完全加权项集维数比率(all-weighted Itemset Dimension Ratio,awIDR),简称项内维数比,即式(15)所示.

$$\text{awIDR}(I_1,I_2)=\frac{k_{12}}{k_1\times k_2}\quad(15)$$

定义 7　基于传统的相关性理论,对于完全加权项集(I_1,I_2),给出其完全加权项集相关性(all-weighted ItemSet Correlation,awISCorr)的计算公式,如式(16)所示.

$$\text{awISCorr}(I_1,I_2)=\frac{\text{awsup}(I_1,I_2)}{\text{awsup}(I_1)\times\text{awsup}(I_2)}\quad(16)$$

根据相关性的性质,在基于项权值变化的完全加权数据挖掘环境下,项集(I_1,I_2)相关性具有如下性质:

性质 1　$\text{awISCorr}(I_1,I_2)>1\Leftrightarrow$项集 I_1和 I_2成正相关;$\text{awISCorr}(I_1,I_2)<1\Leftrightarrow$项集 I_1和 I_2成负相关;$\text{awISCorr}(I_1,I_2)=1\Leftrightarrow$项集 I_1和 I_2无相关性.

性质 2　$\text{awISCorr}(I_1,I_2)>1\Leftrightarrow$①$\text{awISCorr}(I_1,\neg I_2)<1$;②$\text{awISCorr}(\neg I_1,I_2)<1$;③$\text{awISCorr}(\neg I_1,\neg I_2)>1$.$\text{awISCorr}(I_1,I_2)<1\Leftrightarrow$④$\text{awISCorr}(I_1,\neg I_2)>1$;⑤$\text{awISCorr}(\neg I_1,I_2)>1$;⑥$\text{awISCorr}(\neg I_1,\neg I_2)<1$.

证明　证明命题①:$\text{awISCorr}(I_1,I_2)>1\Leftrightarrow$①$\text{awISCorr}(I_1,\neg I_2)<1$.

(1)证明"$\text{awISCorr}(I_1,I_2)>1\Rightarrow\text{awISCorr}(I_1,\neg I_2)<1$"

由式(16)可得,

$$\text{awISCorr}(I_1,I_2)>1\Rightarrow\text{awsup}(I_1,I_2)>\text{awsup}(I_1)\times\text{awsup}(I_2)\quad(17)$$

$$\because\text{awISCorr}(I_1,\neg I_2)=\text{awsup}(I_1,\neg I_2)/(\text{awsup}(I_1)\times\text{awsup}(\neg I_2))\Rightarrow\text{awISCorr}(I_1,\neg I_2)=(\text{awsup}(I_1)-\text{awsup}(I_1,I_2))/(\text{awsup}(I_1)-\text{awsup}(I_1)\times\text{awsup}(I_2))\quad(18)$$

$\because\text{awsup}(I_1)>0,\text{awsup}(I_1)>0,\text{awsup}(I_1,I_2)>0$,由式(17)和(18)得出,$\text{awISCorr}(I_1,\neg I_2)<1$.

$\therefore\text{awISCorr}(I_1,I_2)>1\Rightarrow\text{awISCorr}(I_1,\neg I_2)<1$.

(2)证明"$\text{awISCorr}(I_1,\neg I_2)<1\Rightarrow\text{awISCorr}(I_1,I_2)>1$"

$\because$由式(18)及 $\text{awISCorr}(I_1,\neg I_2)=<1\Rightarrow\text{awsup}(I_1,I_2)>\text{awsup}(I_1)\times\text{awsup}(I_2)\Rightarrow\text{awISCorr}(I_1,I_2)>1$,

$\therefore$ awISCorr(I_1, $\neg I_2$) < 1 $\Rightarrow$awISCorr(I_1, I_2) > 1,

因此,awISCorr(I_1, I_2) > 1 $\Rightarrow$①awISCorr(I_1, $\neg I_2$) < 1.命题②至⑥的证明与上述的类似,证明过程略.证毕.

推论 1 在完全加权数据挖掘环境中,已知项集(I_1, I_2),且 $I_1 \cap I_2 = \varnothing$,①若 $n \times$ awIWR(I_1, I_2) > awIDR(I_1, I_2),则 I_1和 I_2成正相关,能挖掘出 $I_1 \rightarrow I_2$和$\neg I_1 \rightarrow \neg I_2$模式;②若 $n \times$ awIWR(I_1, I_2) < awIDR(I_1, I_2),则 I_1和 I_2成负相关,能挖掘出 $I_1 \rightarrow \neg I_2$和$\neg I_1 \rightarrow I_2$模式;③awIDR(I_1, I_2) > $n \times$ awIWR(I_1, I_2)$\Leftrightarrow$$I_1$和 I_2无相关性.

证明 命题①的证明过程如下:

由式(1)代入式(16)可得到如下式(19).

$$\text{awISCorr}(I_1, I_2) = \frac{n \times k_1 \times k_2 \times w_{12}}{k_{12} \times w_1 \times w_2} = \frac{n \times \text{awIWR}(I_1, I_2)}{\text{awIDR}(I_1, I_2)} \tag{19}$$

$\because$已知 $n \times$ awIWR(I_1, I_2) > awIDR(I_1, I_2), $\therefore$ awISCorr(I_1, I_2) > 1,并由性质 1 和性质 4 可得出命题①成立.

命题②和③的证明过程与命题①的类似,其证明过程略.□

定义 8 设 mc(minimum confidence)为最小置信度阈值,当完全加权项集 I_1和 I_2满足如下 3 个条件,则称关联规则 $I_1 \rightarrow I_2$、$\neg I_1 \rightarrow \neg I_2$、$I_1 \rightarrow \neg I_2$和$\neg I_1 \rightarrow I_2$为有效的完全加权正负关联规则:① I_1和 I_2是频繁项集,$I_1 \cap I_2 = \varnothing$;② $I_1 \rightarrow I_2$、$\neg I_1 \rightarrow \neg I_2$、$I_1 \rightarrow \neg I_2$和$\neg I_1 \rightarrow I_2$的支持度大于或者等于 ms;③ $I_1 \rightarrow I_2$、$\neg I_1 \rightarrow \neg I_2$、$I_1 \rightarrow \neg I_2$和$\neg I_1 \rightarrow I_2$的 awCPIR 值不小于 mc.

4 完全加权项集挖掘算法

4.1 有趣的完全加权频繁项集和负项集剪枝策略

为了发现和剪除那些无趣的项集和规则,兴趣度作为关联模式新度量得到广泛研究和应用[21,22].鉴于此,本文给出了有趣的完全加权频繁项集(Interesting All-Weight Frequent Itemset, IAWFI)和完全加权负项集(Interesting All-Weight Nagetive Itemset, IAWNI)的评判条件,如式(20)和式(21)所示,其中 mi(minimum itemset)为最小兴趣度阈值.因此,有趣的完全加权频繁项集和负项集 I 的剪枝策略是:将不满足 IAWFI(I)条件的频繁项集以及不满足 IAWNI(I)条件的负项集剪除.

$$\begin{aligned}\text{IAWFI}(I) &= \exists I_1, I_2 \subset I: I_1 \cap I_2 \\ &= \varnothing \wedge I_1 \cup I_2 \\ &= I \wedge \text{awsup}(I_1) \geqslant \text{ms} \wedge \text{awsup}(I_2) \\ &\geqslant \text{ms} \wedge (\text{awII}(I_1, I_2) \geqslant \text{mi} \vee \text{awII}(\neg I_1, \neg I_2) \\ &\quad \geqslant \text{mi}) \end{aligned} \tag{20}$$

$$\begin{aligned}\text{IAWNI}(I) &= \exists I_1, I_2 \subset I: I_1 \cap I_2 \\ &= \varnothing \wedge I_1 \cup I_2 \\ &= I \wedge \text{awsup}(I_1) \\ &\geqslant \text{ms} \wedge \text{awsup}(I_2) \\ &\geqslant \text{ms} \wedge (\text{awII}(I_1, \neg I_2) \geqslant \text{mi} \vee \text{awII}(\neg I_1, I_2) \\ &\geqslant \text{mi} \vee \text{awII}(\neg I_1, \neg I_2) \geqslant \text{mi}) \end{aligned} \tag{21}$$

4.2 完全加权项集挖掘算法设计

根据上述剪枝策略,给出完全加权频繁项集和负项集挖掘算法 AWFNIS-Mining(All-Weighted Frequent and Negative Itemsets Mining).

算法 1 AWFNIS-Mining

输入:AWD, ms, mi.

输出:awPIS:完全加权频繁项集集合,awNIS:完全加权负项集集合.

(1)let awPIS = $\varnothing$; awNIS = $\varnothing$;

(2)let $L_1 \leftarrow${在 AWD 中挖掘完全加权频繁 1-项集}; awPIS$\leftarrow$awPIS$\cup$ L_1;

(3)for (i = 2; $L_{i-1} \neq \varnothing$; i + +) do

begin

① $C_i \leftarrow${L_{i-1}进行 Apriori 连接[1]生成候选 i_项集};

②(weight(C_i), awsup(C_i))$\leftarrow${计算 C_i在 AWD 中的权值及其支持度};

③(L_i, N_i)$\leftarrow${从 C_i中挖掘完全加权频繁 i-项集 L_i和负 i-项集 N_i};

④awPIS$\leftarrow$awPIS$\cup$ L_i; awNIS$\leftarrow$awNIS$\cup$ N_i;

end;

(4)forawPIS 集合中每个频繁 i-项集 L_i do

begin

①计算 IAWFI(L_i)值;

②if IAWFI(L_i)值为假 then 从 awPIS 集合中剪除 L_i;

end;

(5)for awNIS 集合中每个负 i-项集 N_i do

begin

①计算 IAWNI(N_i)值;

②if IAWNI(N_i)值为假 then 从 awNIS 集合中剪除 N_i;

end;

(6)输出 awPIS 和 awNIS.

AWFNIS-Mining 算法的时间复杂度为

$$\max\left\{O\left(n \times \sum_{i=1}^{k} |C_i|\right), O\left(\sum_{j=2}^{k-1} |L_{j-1}|^2\right)\right\}$$

其中,n 为 AWD 中的事务记录总数,$|C_i|$、$|L_{j-1}|$分别为候选项集 C_i和频繁项集 L_{j-1}的项目个数,k 为项集维数.

5 基于 SCPIRCI 框架的完全加权正负关联规则挖掘算法

5.1 SCPIRCI:完全加权正负关联规则评价框架

在数据挖掘中,支持度-置信度框架是早期关联模式的评价标准,其缺陷是:无法区分正相关、负相关和不相关的规则,产生矛盾或错误的规则模式.支持度-置信度-相关性框架作为正负关联模式评价标准,有效地避免相互矛盾的模式出现,但还产生无效的、无趣的模式.基于上述问题,在完全加权模式挖掘环境中,构建支持度-CPIR 模型-相关性-兴趣度评价框架,即 SCPIRCI(Support-Conditional Probability Increment Ratio-Correlation-

Interest)框架,将支持度、CPIR 模型、相关性和兴趣度集成,综合对完全加权关联规则进行评价,以减少无趣的和无效的关联模式产生.同时,将同时满足支持度、CPIR 模型和兴趣度要求的关联规则称为有效的完全加权正负关联规则.

5.2 算法设计

算法设计思想是:首先通过上述 AWFNIS-Mining 算法挖掘有趣的频繁项集和负项集,然后,通过项集的项内权值比和维数比简单计算和比较,从频繁项集和负项集挖掘有效的完全加权正负关联规则.具体挖掘过程形式化为 AWPNAR-Mining(All-Weighted Positive and Negative Association Rules Mining)算法.

算法 2　AWPNAR-Mining

输入:AWD,ms,mi,mc.

输出:awPAR:有效的完全加权正关联规则集合;awNAR:有效的完全加权负关联规则集合.

(1)(awPIS,awNIS)←{调用 AWFNIS-Mining 算法产生有趣的频繁项集和负项集};

(2)for awPIS 集合中每个频繁 i-项集 L_i do

begin

if ($I_1, I_2 \subset L_i$: $I_1 \cup I_2 = L_i \wedge I_1 \cap I_2 = \varnothing \wedge \text{awsup}(I_1) \geqslant \text{ms} \wedge \text{awsup}(I_2) \geqslant \text{ms}$) then

begin

①计算 awIWR(I_1, I_2)和 awIDR(I_1, I_2);

②if ($n \times \text{awIWR}(I_1, I_2) > \text{awIDR}(I_1, I_2)$) then

begin

if ($\text{awsup}(I_1, I_2) \geqslant \text{ms}$) then

begin

if ($\text{awCPIR}(I_1 \rightarrow I_2) \geqslant \text{mc}$) then awPAR←awPAR∪{$I_1 \rightarrow I_2$};

if ($\text{awCPIR}(I_2 \rightarrow I_1) \geqslant \text{mc}$) then awPAR←awPAR∪{$I_2 \rightarrow I_1$};

end;

if ($\text{awsup}(\neg I_1, \neg I_2) \geqslant \text{ms}$) then

begin

if ($\text{awCPIR}(\neg I_1 \rightarrow \neg I_2) \geqslant \text{mc}$) then awNAR←awNAR∪{$\neg I_1 \rightarrow \neg I_2$};

if ($\text{awCPIR}(\neg I_2 \rightarrow \neg I_1) \geqslant \text{mc}$) then awNAR←awNAR∪{$\neg I_2 \rightarrow \neg I_1$};

end;

end;

③if ($n \times \text{awIWR}(I_1, I_2) < \text{awIDR}(I_1, I_2)$) then

begin

if ($\text{awsup}(I_1, \neg I_2) \geqslant \text{ms}$) then

begin

if ($\text{awCPIR}(I_1 \rightarrow \neg I_2) \geqslant \text{mc}$) then awNAR←awNAR∪{$I_1 \rightarrow \neg I_2$};

if ($\text{awCPIR}(\neg I_2 \rightarrow I_1) \geqslant \text{mc}$) then awNAR←awNAR∪{$\neg I_2 \rightarrow I_1$};

end;

if ($\text{awsup}(\neg I_1, I_2) \geqslant \text{ms}$) then

begin

if ($\text{awCPIR}(\neg I_1 \rightarrow I_2) \geqslant \text{mc}$) then awNAR←awNAR∪{$\neg I_1 \rightarrow I_2$};

if ($\text{awCPIR}(I_2 \rightarrow \neg I_1) \geqslant \text{mc}$) then awNAR←awNAR∪{$I_2 \rightarrow \neg I_1$};

end;

end;

end;

(3)forawNIS 集合中每个负 i-项集 N_i do

begin

if ($I_1, I_2 \subset N_i$: $I_1 \cup I_2 = N_i \wedge I_1 \cap I_2 = \varnothing \wedge \text{awsup}(I_1) \geqslant \text{ms} \wedge \text{awsup}(I_2) \geqslant \text{ms}$) then

begin

①计算 awIWR(I_1, I_2)及其 awIDR(I_1, I_2);

②if ($n \times \text{awIWR}(I_1, I_2) > \text{awIDR}(I_1, I_2)$) then

begin

if ($\text{awsup}(\neg I_1, \neg I_2) \geqslant \text{ms}$) then

begin

if ($\text{awCPIR}(\neg I_1 \rightarrow \neg I_2) \geqslant \text{mc}$) then awNAR←awNAR∪{$\neg I_1 \rightarrow \neg I_2$};

if ($\text{awCPIR}(\neg I_2 \rightarrow \neg I_1) \geqslant \text{mc}$) then awNAR←awNAR∪{$\neg I_2 \rightarrow \neg I_1$};

end;

end;

③if ($n \times \text{awIWR}(I_1, I_2) < \text{awIDR}(I_1, I_2)$) then

begin

if ($\text{awsup}(I_1, \neg I_2) \geqslant \text{ms}$) then

begin

if ($\text{awCPIR}(I_1 \rightarrow \neg I_2) \geqslant \text{mc}$) then awNAR←awNAR∪{$I_1 \rightarrow \neg I_2$};

if ($\text{awCPIR}(\neg I_2 \rightarrow I_1) \geqslant \text{mc}$) then awNAR←awNAR∪{$\neg I_2 \rightarrow I_1$};

end;

if ($\text{awsup}(\neg I_1, I_2) \geqslant \text{ms}$) then

begin

if ($\text{awCPIR}(\neg I_1 \rightarrow I_2) \geqslant \text{mc}$) then awNAR←awNAR∪{$\neg I_1 \rightarrow I_2$};

if ($\text{awCPIR}(I_2 \rightarrow \neg I_1) \geqslant \text{mc}$) then awNAR←awNAR∪{$I_2 \rightarrow \neg I_1$};

end;

end;

end;

(4)输出 awPAR 和 awNAR.

AWPNAR-Mining 算法的时间复杂度为

$$\max\Big\{O\Big(n \times \sum_{i=1}^{k} |C_i|\Big), O\Big(\sum_{j=2}^{k-1} |L_{j-1}|^2\Big), O\Big(\sum_{i=1}^{n_L} n_{\text{sub}L_i}\Big), O\Big(\sum_{i=1}^{n_N} n_{\text{sub}N_i}\Big)\Big\}$$

其中 n_L 为 awPIS 中频繁项集个数,n_N 为 awNIS 中负项集个数,$n_{\text{sub}L_i}$和 $n_{\text{sub}N_i}$分别为 L_i和 N_i的所有真子集个数.

6 实验与分析

6.1 实验数据

选择北京大学网络实验室提供的测试集 CWT200g(详见 http://www.cwirf.org/)的部分中文语料(12024 篇,编号:CWT200g060412-00750750-23123411)和国外标准测试集 NTCIR-5(详见 http://research.nii.ac.jp/ntcir/permission/ntcir-5/perm-en-CLIR.html)的 Korea_Times2001 英文语料(4936 篇,编号:KT2001_00000-05066)作为实验数据.经过中英文档预处理后,构建基于向量空间模

型的文本数据库.在文档预处理时,中文特征词权值计算公式[19]为:$w=(0.5+0.5\times tf/\max(tf))\times idf$,英文的权值公式[20]为:$w=(1+\ln(tf))\times idf$),中文文档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(Institute of Computing Technology, Chinese Lexical Analysis System),英文文档词干提取程序采用 Porter(见 http://tartarus.org/~martin/PorterStemmer)程序.实验中去掉文档频度(即含有该特征词的文档数量)df 值比较低和比较高的特征词,构建特征词项目库(中文特征词数量:400($1500\leqslant df\leqslant 5838$),英文特征词数量:50($1028\leqslant df\leqslant 2593$)).实验参数如下:ms, mc, mi, ItemNum(所挖掘的项目数量),TRecordNum(事务记录总数),ImaxLength(项集的最大长度,实验中设为 4).

6.2 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[8](记为 PNAR-IMLMS) 以及现有的完全加权词间关联规则挖掘算法 AWARM[16]为实验对比算法,从置信度变化、支持度变化、项目数变化和数据测试集规模变化等方面对算法挖掘性能进行实验对比和分析.实验中,对于对比算法 PNAR-IMLMS,设其 1-项集最小支持度阈值为 $minsup(1)=ms+0.006$,同理,2-项集的为 $minsup(2)=ms+0.004$,3-项集的为 $minsup(3)=ms+0.002$,4-项集的为 $minsup(4)=ms$,$\beta=0.001$.

6.2.1 置信度阈值变化情况下挖掘性能比较

在中文文档测试集 CWT200g 中,设置置信度阈值 mc 为 0.001,0.01,0.1,0.3,0.5,在英文文档测试集 NTCIR-5 中,设置 mc 为 0.01,0.03,0.05,0.07,0.09,0.1 时,本文算法 AWPNAR-Mining 和对比算法在中英文实验文档测试集中挖掘正负关联规则数量总和如表 3 和表 4 所示.

表 3 在 CWT200g 不同置信度下挖掘的正负关联规则数量总和

(ms = 0.03, mi = 0.0002, ItemNum = 50, TRecordNum = 12024)

算法	$A\rightarrow B$	$A\rightarrow \neg B$	$\neg A\rightarrow B$	$\neg A\rightarrow \neg B$
PNAR-CPIR	219694	31549	15678	683597
PNAR-IMLMS	230379	31848	25695	1194624
AWARM	7466	0	0	0
AWPNAR-Mining	6963	1367	602	52185

表 4 在 NTCIR-5 不同置信度下挖掘的正负关联规则数量总和

(ms = 0.07, mi = 0.0002, ItemNum = 50, TRecordNum = 4936)

算法	$A\rightarrow B$	$A\rightarrow \neg B$	$\neg A\rightarrow B$	$\neg A\rightarrow \neg B$
PNAR-CPIR	22501	25415	1602	288674
PNAR-IMLMS	23868	39600	39600	425532
AWARM	15240	0	0	0
AWPNAR-Mining	14321	3192	181	216317

6.2.2 支持度阈值变化情况下挖掘性能比较

在中文文档测试集 CWT200g 中,支持度阈值 ms 设置为 0.03,0.04,0.05,0.06,0.07,0.08,在英文文档测试集 NTCIR-5 中,支持度阈值 ms 设置为 0.08,0.1,0.13,0.15,0.17 时,本文算法 AWPNAR-Mining 和对比算法挖掘候选项集(Candidate Itemset, CI)、频繁项集(Frequent Itemset, FI)、负项集(Negative Itemset, NI))和正负关联规则数量总和如表 5 和表 6 所示.

表 5 在中文测试集不同支持度阈值下挖掘的各类项集和关联规则数量总和

(CWT200g: mc = 0.0002, mi = 0.0002, ItemNum = 50, TRecordNum = 12024)

算法	CI	FI	NI	$A\rightarrow B$	$A\rightarrow \neg B$	$\neg A\rightarrow B$	$\neg A\rightarrow \neg B$
PNAR-CPIR	68884	13242	54653	96086	16786	16786	397748
PNAR-IMLMS	61069	12268	47521	90216	14082	14082	396236
AWARM	4704	754	0	1862	0	0	0
AWPNAR-Mining	8155	736	7205	1758	843	1536	20456

表 6 在英文测试集不同支持度阈值下挖掘的各类项集数量和关联规则数量总和

(NTCIR-5: mc = 0.01, mi = 0.0002, ItemNum = 50, TRecordNum = 4936)

算法	CI	FI	NI	$A\rightarrow B$	$A\rightarrow \neg B$	$\neg A\rightarrow B$	$\neg A\rightarrow \neg B$
PNAR-CPIR	26951	1959	23876	4304	6493	2446	76424
PNAR-IMLMS	23988	1728	21605	3758	7352	7352	72054
AWARM	3386	975	0	1752	0	0	0
AWPNAR-Mining	11692	941	10480	1723	853	444	34981

6.2.3 可扩展性能分析

从项目数量变化和数据测试集规模变化两种情况对算法可扩展性能实验与分析.在项目数量变化和测试集规模分别变化情况下,本文算法在中文测试集 CWT200g 中挖掘频繁项集(FI)、负项集(NI)、关联规则(Association Rule, AR)和负关联规则(Negative Association Rule, NAR)等模式数量变化结果如图 1 ~ 6 所示(ItemNum = 50, TRecordNum = 12024, ms = 0.05, mc = 0.0002, mi = 0.001).

6.2.4 挖掘时间效率性能比较

在支持度阈值变化下,在 CWT200g 中,设置 mc = 0.0002, TRecordNum = 12024, ms 为 0.03, 0.04, 0.05, 0.06, 0.07, 0.08 和 0.09,在 NTCIR-5 中,设置 mc = 0.01, TRecordNum = 4936, ms 为 0.08, 0.1, 0.13, 0.15 和 0.17;

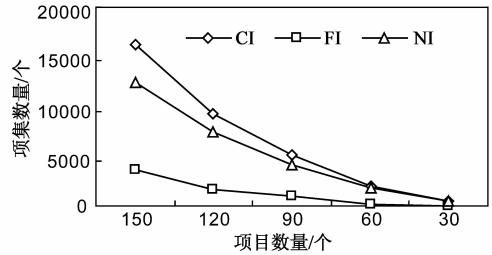

图1 不同项目数的候选、频繁和负项集数量变化

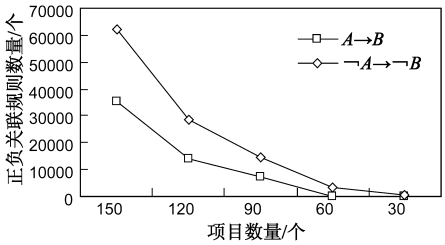

图2　不同项目数的正负关联规则数量变化

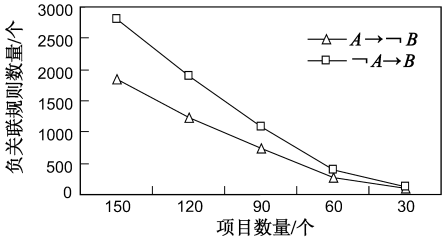

图3　不同项目数的负关联规则数量变化

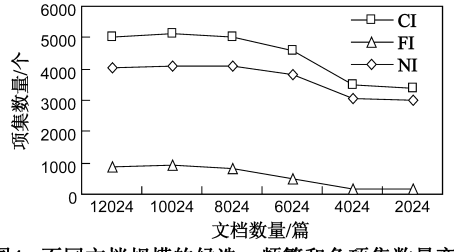

图4　不同文档规模的候选、频繁和负项集数量变化

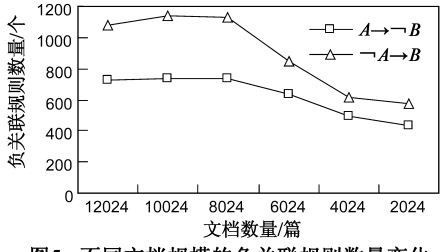

图5　不同文档规模的负关联规则数量变化

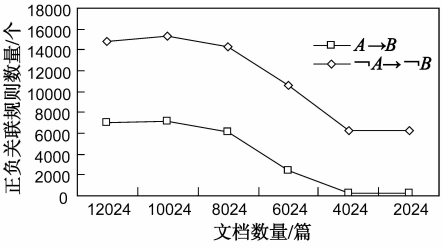

图6　不同文档规模的正负关联规则数量变化

在置信度阈值变化情况下,在 CWT200g 中,设置 ms = 0.03,TRecordNum = 12024,mc 为 0.001,0.01,0.1,0.3,0.5,0.7 和 0.9,在 NTCIR-5 中,设置 ms = 0.07,TRecordNum = 4936,mc 为 0.01,0.03,0.05,0.07 和 0.09,统计 3 种算法在中英文测试集的挖掘时间总和,其结果如表 7 至表 8 所示(mi = 0.0002,ItemNum = 50).

表 7　不同支持度下挖掘项集和关联规则的时间总和(s)

数据集	PNAR-CPIR	PNAR-IMLMS	AWPNAR-Mining
CWT200g	29049	27222	1716
NTCIR-5	5906	4988	2733

表 8　不同置信度下挖掘正负关联规则的时间总和(s)

数据集	PNAR-CPIR	PNAR-IMLMS	AWPNAR-Mining
CWT200g	58801	70536	2681
NTCIR-5	17573	19034	12925

6.2.5　项集的剪枝性能分析

在中文测试集 CWT200g 中,本文算法 AWPNAR-Mining 在支持度阈值变化和兴趣度阈值变化两种情况下进行的剪枝性能实验,结果如表 9 和表 10 所示(mc = 0.0002,ItemNum = 50,TRecordNum = 12024).

表 9　支持度变化下项集的剪枝结果(mi = 0.07)

ms	不剪枝		剪枝	
	FI	NI	FI	NI
0.03	389	2594	279	1853
0.04	141	1445	114	659
0.05	83	1213	74	410
0.06	64	1168	61	397
0.07	43	711	42	285
0.08	26	278	26	150
0.09	4	52	4	50
合计	750	7461	600	3804

表 10　兴趣度变化下项集的剪枝结果(ms = 0.01)

ms	频繁项集		负项集	
	数量(个)	减幅(%)	数量(个)	减幅(%)
0	8308	0	30070	0
0.050	8276	0.39	29940	0.43
0.070	7482	9.94	27748	7.72
0.075	6340	23.69	25016	16.81
0.080	1765	78.76	9047	69.91
0.085	190	97.71	2750	90.85
0.090	77	99.07	157	99.48
0.100	50	99.40	0	0

6.2.6　关联模式实例分析

在中文文本数据集 CWT200g 中,取 5 个特征词项目(即,部门(df:1898),采用(df:1825),参加(df:1668),参与(df:1512),产品(df:2284))进行挖掘,实验参数为:ms = 0.01,mi = 0.0002,ItemNum = 5,TRecordNum = 12024,挖掘所得的各类 3-项集结果实例如表 11 所示.

从表 11 可以看出,本文算法挖掘的频繁 3-项集和负 3-项集数量比对比算法 PNAR-CPIR、PNAR-IMLMS 挖

掘的少,所得的项集更合理、更接近实际情况,例如,3-项集"{部门,参加,参与}"在上述对比算法挖掘的结果中是频繁项集,这是不合理的,应是无效和无趣的项集模式,而在本文算法挖掘的结果中是负项集,这是合理的,应该是个真实的、有效的项集模式.其原因分析如下:由于"参加"和"参与"是近义词,在一句话或者一段话中应该很少同时出现,属于负相关关联,因此,项集"{部门,参加,参与}"是负项集更合理.AWARM 算法和本文算法是基于项权值变化的模式挖掘算法,其频繁项集结果相同,但 AWARM 算法没能挖掘出负项集.

表 11 4 种算法在中文数据集 CWT200g 挖掘的 3-项集实例

算法	项集	数量	项集实例
PNAR-CPIR	CI	10	{部门,参与,产品},{部门,参加,参与},{部门,参加,产品},{部门,采用,参加},{部门,采用,参与},{部门,采用,产品},{采用,参与,产品,},{采用,参加,参与},{采用,参加,产品},{参加,参与,产品}
	FI	9	{部门,参与,产品},{部门,参加,参与},{部门,参加,产品},{部门,采用,参加},{部门,采用,参与},{部门,采用,产品},{采用,参与,产品,},{采用,参加,参与},{参加,参与,产品}
	NI	1	{采用,参加,产品}
PNAR-IMLMS (minsup(3) = 0.012)	CI	10	同 PNAR-CPIR 算法的
	FI	7	{部门,参与,产品},{部门,参加,参与},{部门,采用,参加},{部门,采用,参与},{部门,采用,产品},{采用,参与,产品,},{采用,参加,参与},
	NI	3	{部门,参加,产品},{采用,参加,产品},{参加,参与,产品}
AWARM	CI	7	{部门,参与,产品},{部门,参加,参与},{部门,采用,参加},{部门,采用,参与},{部门,采用,产品},{采用,参加,参与},{参加,参与,产品}
	FI	3	{部门,采用,参与},{部门,采用,产品},{部门,参与,产品}
AWPNAR-Mining	CI	10	同 PNAR-CPIR 算法的
	FI	3	{部门,采用,参与},{部门,采用,产品},{部门,参与,产品}
	NI	7	{部门,参加,参与},{部门,参加,产品},{部门,采用,参加},{采用,参与,产品},{采用,参加,参与},{采用,参加,产品},{参加,参与,产品}

6.2.7 实验结果分析

上述实验结果表明,与 3 种对比算法比较,在中英文标准数据测试集中,本文算法具有如下特点:

(1)无论是支持度阈值变化或者置信度阈值变化,本文算法 AWPNAR-Mining 挖掘的候选项集、频繁项集、负项集和正负关联规则数量都比对比算法 PNAR-CPIR、PNAR-IMLMS 挖掘的少,降幅比较大.如表 5 和表 7 所示,在中文数据集的实验结果表明,本文算法挖掘的候选项集数量比 PNAR-CPIR、PNAR-IMLMS 挖掘的分别减少 88.16% 和 86.65%,挖掘时间比 PNAR-CPIR、PNAR-IMLMS 挖掘的分别减少 94.09% 和 93.70%.

(2)本文算法的挖掘时间比对比算法 PNAR-CPIR、PNAR-IMLMS 的少,减幅较大,表明本文算法挖掘效率得到极大地提高.

(3)随着项目数量或者事务文档数量增多,本文算法挖掘各类项集的数量逐渐增多,表现出良好的可扩展性.

(4)本文算法具有良好的剪枝性能,那些无趣的和无效的频繁项集和负项集得到了有效地排除.

主要原因分析如下:对比算法 PNAR-CPIR、PNAR-IMLMS 是基于项目频度挖掘的无加权正负关联规则挖掘算法,没有考虑项权值,忽略完全加权数据具有项权值变化的固有特点,产生很多无效的和虚假的项集,使得项集数量增多,其挖掘效率大大减低.对比算法 AWARM 虽然是完全加权模式挖掘算法,但不能挖掘负关联模式.本文算法 AWPNAR-Mining 属于基于项权值变化的完全加权正负关联规则挖掘算法,有效地克服了对比算法的缺陷,重视项权值依赖于事务的特点,所挖掘的正负关联规则模式更合理、更接近实际,同时,采用了新的剪枝策略,使得无效和无趣的模式数量大幅度减少,提高了挖掘效率.

7 结论

基于项权值变化的挖掘技术在文本挖掘、信息检索等领域具有重要的理论价值和广阔的应用前景,因此,深入研究基于项权值变化的正负关联模式挖掘技术是必要的.本文对基于项权值固定的和基于项权值变化的数据模型进行比较性研究,提出一种基于项权值变化的完全加权正负关联规则挖掘方法,解决了基于项权值变化的负关联规则挖掘技术问题.该方法构建一种基于项权值变化的完全加权关联模式评价框架 SCPIRCI 及其项集剪枝策略,设计基于 SCPIRCI 评价框架的完全加权正负关联规则挖掘算法,考虑了项权值依赖于事务记录的特点,采用新的项集剪枝方法和模式评价框架挖掘有趣的频繁项集和负项集,通过项集的项内权值比和维数比的简单计算和比较,从频繁项集和负项集中挖掘更合理、更能接近实际情况的完全加权正负关联规则模式.实验结果表明本文算法具有良好的效果.下一步的研究重点是:探索将该成果运用于信息检索查询扩展领域,以提高信息检索查询性能.

参考文献

[1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[A]. Proceedings of 1993 ACM-SIGMOD International Conference on Management of Data(SIGMOD. 93)[C]. New York, USA: ACM Press, 1993. 207 – 216.

[2] 宋威,李晋宏,徐章艳,等.一种新的频繁项集精简表示方法及其挖掘算法的研究[J].计算机研究与发展,2010,47(2):277 – 285.
Song Wei, Li Jin-hong, Xu Zhang-yan, et al. Research on a new concise representation of frequent itemset and its mining algorithm[J]. Journal of Computer Research and Development, 2010,47(2):277 – 285.(in Chinese)

[3] Nnrmadha D, Naveensundar G, Geetha S. Anovel approach to prune mined association rules in large databases[A]. Proceeding of 2011 3rd International Conference on Electronics Computer Technology(ICECT)[C]. Washington, USA: IEEE Computer Society Press, 2011. 409 – 413.

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

[5] 耿生玲,李永明,刘震.关联规则挖掘的软集包含度方法[J].电子学报,2013,41(4):804 – 809.
Geng Sheng-ling, Li Yong-ming, Liu Zhen. A approach to association rules mining using inclusion degree of soft sets[J]. Acta Electronic Sinica, 2013,41(4):804 – 809.(in Chinese)

[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 D, Zhang C Q, Zhang S C. Efficient mining of both positive and negative association rules[J]. ACM Transactions on Information Systems, 2004, 22(3):381 – 405.

[8] Swesi I M A O, Bakar A A, Kadir A S A. Mining positive and negative association rules from interesting frequent and infrequent itemsets[A]. Proceedings of 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012)[C]. Washington, USA: IEEE Computer Society Press, 2012. 650 – 655.

[9] Bhargava R, Lade S. Effectivepositive negative association rule mining using improved frequent pattern tree[J]. International Journal of Advanced Research in Computer Science and Software Engineering, 2013, 3(4):193 – 199.

[10] Cai C H, Da A, Fu W C, et al. Mining association rules with weighted items[A]. Proceedings of IEEE International Database Engineering and Application Symposiums[C]. Washington, USA: IEEE Computer Society Press, 1998. 68 – 77.

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

[12] Pears R, Koh Y S. Weighted association rule mining using particle swarm optimization[A]. Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) Workshops[C]. Berlin, Germany: Pringer-Verlag, 2012. 327 – 338.

[13] Tan J. Weighted association rules mining algorithm research[J]. Applied Mechanics and Materials, 2013, 241-244: 1598 – 1601.

[14] Zhao Y Y, Jiang H, Geng R, et al. Mining weighted negative association rules based on correlation from infrequent items[A]. Proceedings of the 2009 International Conference on Advanced Computer Control[C]. Los Alamitos, California, USA: IEEE Computer Society Press, 2009. 270 – 273.

[15] 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(13):208 – 211.
Tan Yi-hong, Lin Ya-ping. Mining all-weighted association rules from vector space model[J]. Computer Engineering and Applications, 2003(13):208 – 211.(in Chinese)

[16] 黄名选,严小卫,张师超.基于文本库的完全加权词间关联规则挖掘算法[J].广西师范大学学报(自然科学版),2007,25(4):24 – 27.
Huang Ming-xuan, Yan Xiao-wei, Zhang Shi-chao. Algorithm of item-all-weighted association rules mining between terms from text database[J]. Journal of Guangxi Normal University (Natural Science Edition), 2007, 25(4):24 – 27.(in Chinese)

[17] 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 Ming-xuan, Yan Xiao-wei, Zhang Shi-chao. Query expansion of pseudo relevance feedback based on matrix-weighted association rules mining[J]. Journal of Software, 2009, 20(7):1854 – 1865.(in Chinese)

[18] 刘远超,王晓龙,徐志明,等.基于粗集理论的中文关键词短语构成规则挖掘[J].电子学报,2007,35(2):371 – 374.
Liu Yuan-chao, Wang Xiao-long, Xu Zhi-ming, et al. Mining construction rules of chinese keyphrase based on rough set theory[J]. Acta Electronic Sinica, 2007, 35(2):371 – 374.(in Chinese)

[19] Grossman D A, Frieder O. 信息检索:算法与启发式方法[M].北京:人民邮电出版社,2010. 12 – 13.

[20] Büttcher S, Clarke C L A, Cormack G V. 信息检索实现和评价搜索引擎[M].北京:机械工业出版社,2012. 40 – 41.

[21] 程继华,郭建生,施鹏飞.挖掘所关注规则的多策略方法研究[J].计算机学报,2000,23(1):47 – 51.
Cheng Ji-hua, Guo Jian-sheng, Shi Peng-fei. Multi-strategy approach to mining interesting rules[J]. Chinese Journal of

Computers, 2000, 23(1): 47 – 51. (in Chinese)
[22] 周欣,沙朝锋,朱扬勇,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.
Zhou Xin, Sha Chao-feng, Zhu Yang-yong, et al. Interest measure- – another threshold in association rules[J]. Journal of Computer Research & Development, 2000, 37(5): 627 – 633. (in Chinese)

作者简介

周秀梅 女,1972 年 6 月出生于广西上林县,副教授,主要研究方向为数据挖掘.
E-mail: xm2037@163.com

黄名选(通信作者) 男,1966 年 8 月出生于广西乐业县,工学硕士,广西财经学院计算机系教授,主要研究方向为数据挖掘、信息检索,广西科技项目评估咨询专家,主持或参与国家自然科学基金项目 2 项,主持广西自然科学基金项目 1 项,主持广西高等学校科研项目 3 项,入选 2011 年度"广西高校优秀人才资助计划"资助人选,发表学术论文 50 余篇,其中,中文核心期刊论文 30 余篇,被期刊 EI 索引 2 篇,被 ISTP 索引 1 篇,申请发明专利 11 项.
E-mail: huangmx@mailbox.gxnu.edu.cn