# Interestingness Measures for Association Patterns : A Perspective *

Pang-Ning Tan
Department of Computer Science
University of Minnesota
Minneapolis, MN 55455
ptan@cs.umn.edu

Vipin Kumar
Department of Computer Science
University of Minnesota
Minneapolis, MN 55455
kumar@cs.umn.edu

## ABSTRACT

Association rules are valuable patterns because they offer useful insight into the types of dependencies that exist between attributes of a data set. Due to the completeness nature of algorithms such as Apriori, the number of patterns extracted are often very large. Therefore, there is a need to prune or rank the discovered patterns according to their degree of interestingness. In this paper, we will examine the various interestingness measures proposed in statistics, machine learning and data mining literature. We will compare these measures and investigate how close they reflect the statistical notion of correlation. We will show that support-based pruning, which is often used in association rule discovery, is appropriate because it removes mostly uncorrelated and negatively correlated patterns. Our experimental results verified that many of the intuitive measures (such as Piatetsky-Shapiro's rule-interest, confidence, laplace, entropy gain, etc.) are very similar in nature to correlation coefficient (in the region of support values typically encountered in practice). Finally, we will introduce a new metric, called the $IS$ measure, and show that it is highly linear with respect to correlation coefficient for many interesting association patterns.

## 1. INTRODUCTION

Association rules [2, 1] are valuable patterns that can be derived from large databases. Conceptually, an association rule indicates that the presence of a set of items (itemset) in a transaction would imply the occurence of other items in the same transaction. The association rule discovery problem is often decomposed into two separate tasks : (1) to discover all itemsets having support above a user-defined threshold, and (2) to generate rules from the frequent itemsets. The first task can be very expensive, because it may require a lot of I/O operations. Over the years, many algorithms have been developed to efficiently generate the frequent itemsets [3, 14, 8].

The rule generation task is less I/O intensive. However, there are two major problems with association rule generation : (1) too many rules are generated (rule quantity problem), and (2) not all of the rules are interesting (rule quality problem). Both problems are not entirely independent. For example, knowledge about the quality of a rule can be used to reduce the number of rules presented to an analyst.

There has been various research effort aimed at mitigating both problems. The rule quantity problem can be handled by pruning or summarizing the discovered rules. Toivonen et al.[19] proposed the idea of using structural rule covers to remove redundant rules and clustering as a means for grouping together related rule covers. Liu et al. [12] used the standard $\chi^2$ test to prune insignificant rules and introduced the concept of direction setting rules to summarize the patterns. Other researchers such as Srikant et al. [18] and Ng et al. [13] have used the constraints provided by a user to limit the number of rules that are generated.

Solution to the rule quality problem relies on specification of an interestingness measure to represent the novelty, utility or significance of a pattern. By ordering the discovered rules according to their degree of interestingness, highly-ranked rules can be presented to the analyst. Some of these measures are applicable to itemsets as well as rules. [1] In such cases, they can be incorporated into the itemset generation step to remove uninteresting itemsets.

Support and confidence are used in the original formulation of association rule discovery problem. Support is necessary because it represents the statistical significance of a pattern. From the marketing perspective, support of an itemset in retail sales data justifies the feasibility of promoting the items together. Support is also good for pruning the search space since it possesses a nice downward closure (anti-monotonicity) property (Figure 1). This property states that if a set of items $C$ appears in $t$ transactions, then

*This work was supported by NSF ACI-9982274 and by Army High Performance Computing Research Center contract number DAAH04-95-C-0008. Access to computing facilities was provided by AHPCRC, Minnesota Supercomputer Institute.

[1] We will use the term association pattern to refer to both an association rule and the itemset from which the rule is generated.

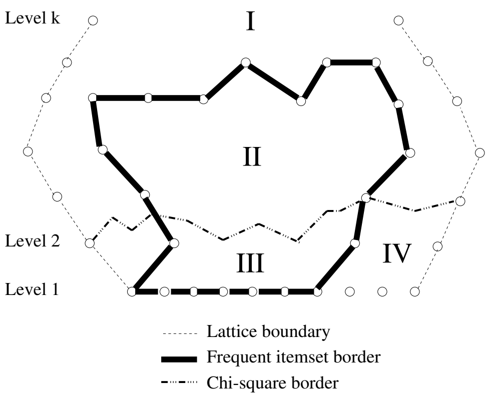


**Figure 1: Lattice structure for all itemsets. This structure can be divided into several regions : I. Infrequent and dependent itemsets, II. frequent and dependent itemsets, III. frequent and independent itemsets, IV. infrequent and independent itemsets.**

any subset of $C$ must occur in at least $t$ transactions. As a result, if an itemset $C$ does not meet the minimum support requirement, then we can ignore all supersets of $C$ from consideration. However, support alone may not serve as a reliable interestingness measure. For example, rules with high support quite often correspond to obvious knowledge about the domain. The rule $Bread \Longrightarrow Milk$, for instance, may not be informative despite having a large support value. In Fig. 1, any itemsets that lie outside the frequent itemset border can be declared as uninteresting. However, one may still have to face the problem of combinatorial explosion due to the large number of rules that can be potentially generated.

Confidence measures the conditional probability of events associated with a particular rule. For example, if a rule $X \longrightarrow Y$ has confidence $c$, this means that $c\%$ of all transactions that contain $X$ will also contain $Y$. Unfortunately, the confidence measure can be misleading in many practical situations, as shown by Brin et al. in [6, 17]. [6, 17] offerred an alternative to evaluate the significance of association patterns using $\chi^2$ test. This test is desirable because it will rule out itemsets that occur by chance. [6, 17] also showed that the $\chi^2$ statistic has an upward closure (monotonicity) property i.e. if an itemset $C$ passes the $\chi^2$ test, so will every superset of $C$. This property will allow us to look for a border between dependent and independent itemsets (Fig. 1). However, we will argue that the alternative proposed in [6, 17] may still be unsatisfactory.

This paper intends to follow-up on the earlier work done by Brin et al in [6]. The main contributions of this paper are as follows :

1. We investigate the possibility of using various measures from statistics, machine learning and data mining literature to rank the association patterns.
2. We show that support-based pruning is useful for removing uncorrelated and negatively correlated itemsets.
3. We combine support-based pruning with $\chi^2$ pruning to reduce the complexity of mining interesting association patterns. Specifically, we examine the applicability of various interestingness measures to region II of Fig. 1.
4. We introduce a new measure, called the $IS$ measure, which takes into account both the interestingness and support of a pattern.
5. We evaluate how well the various interestingness measures can capture the notion of statistical correlation. In fact, our empirical results show that many of these measures are capable of representing statistical correlation within certain range of support values.

## 2. STATISTICAL MEASURES OF DEPENDENCY

In this section, we will present several statistical methods for measuring the dependencies between variables. Our focus will be primarily on pairs of dichotomous variables, even though some of the techniques described here can be extended to more than two variables. In general, the relationship between two binary variables, $A$ and $B$, can be summarized in a $2 \times 2$ contingency table as shown in Table 1.

For comparison purposes, we have generated an artificial dataset that contains 10000 random samples. Each sample is a 4-tuple $(f_{11}, f_{10}, f_{01}, f_{00})$, subjected to the following constraints : $f_{11}/N < 1$, $f_{10}/N < 1$, $f_{01}/N < 1$ and $f_{11} + f_{10} + f_{01} \leq N$. We can think of each sample as a realization of the contingency table for a pair of items (itempair) in the overall dataset.

### 2.1 $\chi^2$ test

The $\chi^2$ test can be used to test the hypothesis that items in a pattern are independent of each other. For the $2 \times 2$ contingency table shown in Table 1, its $\chi^2$ value is given by:

$$\chi^2 = \frac{N(f_{11}f_{00} - f_{01}f_{10})^2}{f_{1+}f_{0+}f_{+1}f_{+0}} \quad (1)$$

The larger the $\chi^2$ value, the more evidence we have to reject the independence hypothesis. [6, 17] have used this test to find both positive and negatively correlated association patterns. They also showed that $\chi^2$ is upward closed, a property

| | $B$ | $\overline{B}$ | |
|---|---|---|---|
| $A$ | $f_{11}$ | $f_{10}$ | $f_{1+}$ |
| $\overline{A}$ | $f_{01}$ | $f_{00}$ | $f_{0+}$ |
| | $f_{+1}$ | $f_{+0}$ | $N$ |

**Table 1: A $2 \times 2$ contingency table for binary variables.**

that can be exploited to prune the exponential search space. [6] have also devised an efficient algorithm to search for a border between dependent and independent itemsets. However, the $\chi^2$ test alone may not be the ultimate answer due to the following reasons:

1. As stated in [6], $\chi^2$ does not tell us the strength of correlation between items in an association pattern. Instead, it will only help us to decide whether items in the pattern are independent of each other. Thus, it cannot be used for ranking purposes.

2. The upward closure property of $\chi^2$ ensures that all itemsets above the $\chi^2$ border are statistically dependent. In reality, some itemsets above the $\chi^2$ border will be more interesting than others. Therefore, just knowing the border alone is insufficient.

3. The $\chi^2$ statistic depends on the total number of transactions. On the other hand, the $\chi^2$ cutoff value depends only on the degrees of freedom of the attributes (which is 1 for binary attributes) and the significance level desired. For example, the rejection region for binary attributes at 0.05 significance level is 3.84. When the number of transactions are large, the cutoff value can be exceeded by a very large number of itemsets.

## 2.2 Measures of Association

For ranking purposes, we need an explicit measure of variable dependencies. We will present two such statistical measures in this section. They are Pearson's correlation coefficient and Goodman and Kruskal's $\lambda$ coefficient [11]. Other measures include Yule's $Q$ and $Y$ coefficients, uncertainty coefficients, Cramer's contingency coefficients, odds ratio, etc. [16, 11, 20].

### 2.2.1 *Correlation coefficient*

Correlation coefficient measures the degree of linear dependency between a pair of random variables. Theoretically, it is defined as the covariance between two variables, divided by their standard deviations ($\sigma$) :

$$\rho_{AB} = \frac{\mathrm{Cov}(A,B)}{\sigma_A \sigma_B} \ . \tag{2}$$

where $\mathrm{Cov}(A,B) = E(AB) - E(A)E(B)$ and $E(\cdot)$ is the expected value. The range of $\rho_{AB}$ is between $-1$ and $+1$. If the two variables are independent, then $\rho_{AB} = 0$. However, the converse is not necessarily true. It is possible that $\rho_{AB} = 0$ when the variables have strong non-linear dependencies. Fortunately, such a problem does not exist for binary variables.

For binary variables, $\sigma_A = \sqrt{P(A)(1-P(A))}$ where $P(A) = f_{1+}/N$. The correlation coefficient between $A$ and $B$ can be written as

$$\rho_{AB} = \frac{Nf_{11} - f_{1+} \cdot f_{+1}}{\sqrt{f_{1+}f_{0+}f_{+1}f_{+0}}} = \frac{f_{11}f_{00} - f_{10}f_{01}}{\sqrt{f_{1+}f_{0+}f_{+1}f_{+0}}} \ . \tag{3}$$

The above equation is obtained assuming that the contingency table is constructed using data from the entire population. For finite samples, the above equation is equivalent to Pearson's $\phi$-coefficient. For convenience, we will use the term correlation coefficient and $\phi$-coefficient interchangeably for the rest of the paper. Also, for binary variables, the $\phi$-coefficient is closely related to the $\chi^2$ statistic. Upon comparing equation 3 with equation 1, we would obtain $\phi^2 = \chi^2/N$.

A greater concern is that a large positive correlation coefficient may indicate either $A$ and $B$ are highly dependent (i.e. $f_{11}$ is large) or $\overline{A}$ and $\overline{B}$ are highly dependent (i.e. $f_{00}$ is large). The two cases can be distinguished only if the joint support of $(A,B)$ (i.e. $f_{11}$) is taken into consideration.

### 2.2.2 *$\lambda$ Coefficient*

This coefficient was suggested based upon the following idea : if two variables are highly dependent, then the error in predicting the value of one of the variables would be smaller whenever the value of the other variable is known. For example, consider the dependencies between $A$ and $B$. If no other information is available, the best guess we can make about the value of $A$ is $\hat{A} = \arg(\max_k P(A_k))$. The error $\epsilon_A$ in making this guess is $P(\epsilon_A) = 1 - P(\hat{A}) = 1 - \max_k P(A_k)$.

Now, suppose we observe $B = B_1$. With this new information, the best estimate of $A$ is the value that maximizes the conditional probability $\hat{A} = \arg(\max_k P(A_k|B_1))$. The error associated with this estimator is $P(\epsilon_A|B_1) = 1 - \max_k P(A_k|B_1)$. The average prediction error for $A$ given $B$ can be computed by averaging over the entire range of $B$ values :

$$\begin{aligned} P(\epsilon_A|B) &= P(\epsilon_A|B_1)P(B_1) + \cdots + P(\epsilon_A|B_m)P(B_m) \\ &= (1 - \max_k P(A_k|B_1))P(B_1) + \cdots \\ &\quad + (1 - \max_k P(A_k|B_m))P(B_m) \\ &= 1 - \sum_j \max_k P(A_k, B_j) \ . \end{aligned} \tag{4}$$

Goodman and Kruskal defines the index of predictive association for $A$ given $B$ as

$$\lambda_A = \frac{P(\epsilon_A) - P(\epsilon_A|B)}{P(\epsilon_A)} = \frac{\sum_j \max_k f_{jk} - \max_k f_{+k}}{N - \max_k f_{+k}} \tag{5}$$

This equation can be used as an interestingness measure for the rule $B \longrightarrow A$. For an itemset $\{A, B\}$ , we can use the symmetric version of this coefficient :

$$\lambda_{AB} = \frac{\sum_j \max_k f_{jk} + \sum_k \max_j f_{jk} - \max_k f_{+k} - \max_j f_{j+}}{2N - \max_k f_{+k} - \max_j f_{j+}} \quad (6)$$

# 3. MEASURES OF DEPENDENCY FROM DATA MINING

In recent years, various interestingness measures have been proposed to measure the significance of patterns derived using machine learning and data mining techniques. Many of these measures can be adapted to association patterns.

## 3.1 Support and Confidence

As previously mentioned, support is necessary because it measures the statistical significance of a pattern. Since the choice of an appropriate support threshold can be ad-hoc, we need to ensure that support-based pruning will not remove many of the interesting patterns. In this paper, we assume that only positively correlated itemsets are of interest to a data analyst. This is a valid assumption in datasets where the presence of an item in a transaction is more significant than its absence. This appears to be true for a large variety of datasets. Figures 2 and 3 show the effect of applying various support thresholds on our artificial dataset. The first graph in both figures represent the histograms of $\phi$-coefficients for every itempairs in the dataset. These histograms appear to be very similar to a Gaussian distribution. The rest of the histograms show the itempairs that are removed when various support thresholds are imposed.

Figure 2 shows that by imposing an upper bound on support, one may end up pruning uncorrelated, positively correlated and negatively correlated itempairs in equal proportions. In contrast, pruning with minimum support will remove mostly uncorrelated or negatively correlated itempairs (Fig. 3). This result makes sense because itempairs with low support tend to have large values of $f_{10}$, $f_{01}$ or $f_{00}$. This often corresponds to uncorrelated or negatively correlated itemsets. In addition, the positively correlated itemsets that are removed are those that have large values in $f_{00}$. Hence, minimum support-based pruning is a good strategy if we are only interested in positively correlated association rules.

Confidence was initially proposed to measure the strength of an association rule. However, [6] showed that it may produce counter-intuitive results especially when strong negative correlations are present.

Consider the $2 \times 2$ table shown in Table 2. It summarized the purchase of two brands of operating systems at a retail store within a certain time period. Suppose the support and confidence thresholds were set at 5% and 50% respectively. The association rule $Linux \longrightarrow WindowsNT$ would have a 20% support and 67% confidence. Thus, it will pass both threshold conditions and eventually declared to be interesting. However, this information can be misleading. The prior probability that a customer purchases Windows NT is 80%. Once we know that the customer had purchased Linux, the conditional probability that he or she would buy Windows NT reduces to 75%. Hence, the high confidence of the rule $Linux \longrightarrow WindowsNT$ is misleading.

Another confidence-like measure is the laplace function, which is often used to measure the accuracy of classification rules [10]:

$$\text{laplace} = \frac{\sigma(A \cup B) + 1}{\sigma(A) + 2} \quad (7)$$

where $\sigma(A)$ denotes the number of transactions that contain $A$.

## 3.2 Interest and $IS$ Measure

Interest factor is another widely used measure for association patterns [6, 17, 5, 9]. This metric is defined to be the ratio between the joint probability of two variables with respect to their expected probabilities under the independence assumption.

$$I(A,B) = \frac{P(A,B)}{P(A)P(B)} = \frac{f_{11}N}{f_{1+}f_{+1}} \quad (8)$$

The interest factor is a non-negative real number; with a value of 1 corresponding to statistical independence.

The interest factor, $I(A,B)$, is closely related to the $\phi$ coefficient. If we re-arrange equation 3, we can obtain the following :

$$\phi = \frac{\frac{Nf_{11}}{f_{1+}f_{+1}} - 1 \cdot f_{1+}f_{+1}}{\sqrt{f_{1+}f_{0+}f_{+1}f_{+0}}} = \frac{(I-1)\cdot\sqrt{f_{1+}f_{+1}}}{\sqrt{f_{0+}f_{+0}}} \quad (9)$$

Consider the region of low support values, i.e. $\frac{f_{1+}}{N} \ll 1$ and $\frac{f_{+1}}{N} \ll 1$. Both $\frac{f_{0+}}{N}$ and $\frac{f_{+0}}{N}$ will be close to 1. If the items are highly correlated, then $I \gg 1$. In this region of approximation, equation 9 becomes :

$$\phi \approx I\sqrt{\frac{f_{1+}f_{+1}}{N^2}} = \frac{Nf_{11}}{f_{1+}f_{+1}}\sqrt{\frac{f_{1+}f_{+1}}{N^2}} = \sqrt{\frac{Nf_{11}}{f_{1+}f_{+1}}\cdot\frac{f_{11}}{N}}$$
$$= \sqrt{I \times \frac{f_{11}}{N}} \quad (10)$$

This suggests that a good interestingness measure, derivable from statistical correlation, in the region of low support and high interest values is :

$$IS = \sqrt{I \times \frac{f_{11}}{N}} = \sqrt{\frac{P(A,B)\ P(A,B)}{P(A)P(B)}} \quad (11)$$

$IS$ has many desirable properties as an interestingness measure. First of all, it is a product of two important quantities, interest factor and support. This measure takes into account both the interestingness and support aspects of a pattern. Secondly, for binary pairs of variables, $IS$ can be expressed as the geometric mean of confidence for rules that can be generated from the itempair i.e.

$$IS = \sqrt{Conf(A \longrightarrow B) \times Conf(B \longrightarrow A)}.$$

Another useful interpretation of this measure is as the cosine angle between two vectors, i.e. $IS = P(A,B)/\sqrt{P(A)P(B)}$.

Figure 4 shows the relationship between $IS$ and $\phi$ using the artificial dataset. Note the high linearity exhibited by the $IS$ measure, agreeing with the theoretical arguments above. We have also repeated our experiments using real-world datasets. The first dataset is a subset of Reuters

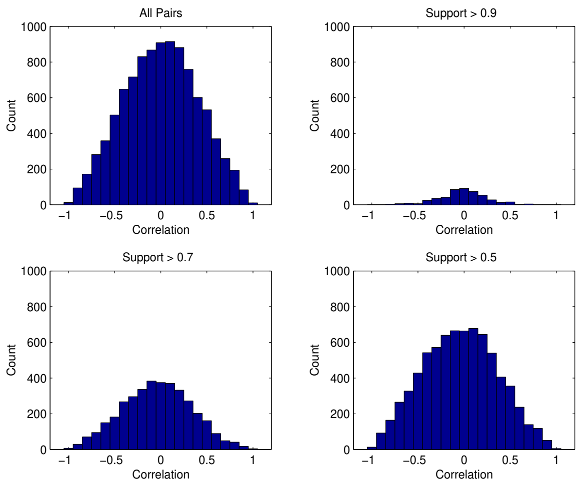


**Figure 2: Number of itempairs removed by applying upper support threshold.**

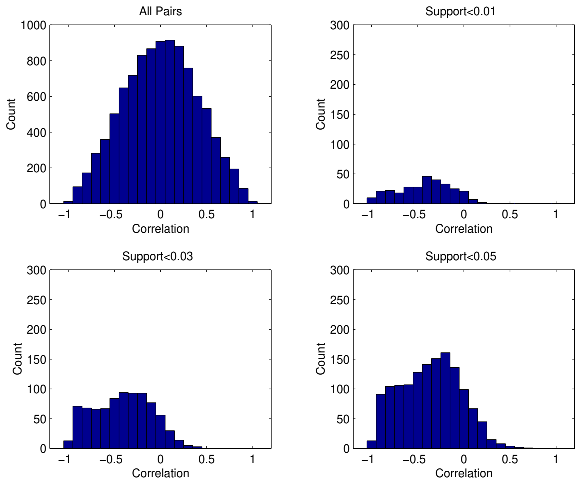


**Figure 3: Number of itempairs removed by applying lower support threshold.**

| | $WindowsNT$ | $\overline{WindowsNT}$ | |
|---|---|---|---|
| $Linux$ | 20 | 10 | 30 |
| $\overline{Linux}$ | 60 | 10 | 70 |
| | 80 | 20 | 100 |

**Table 2: A $2 \times 2$ contingency table example.**

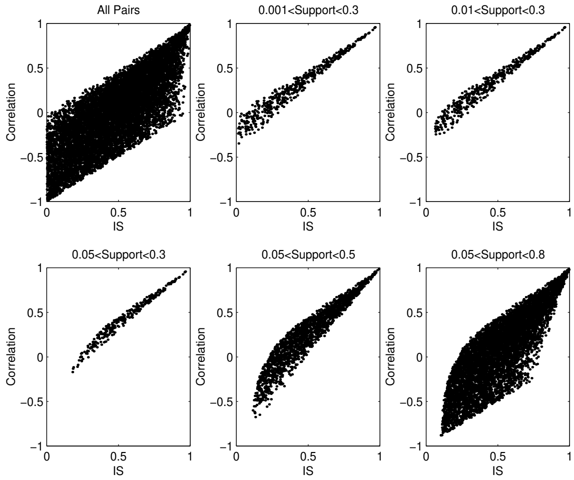


**Figure 4: Correlation coefficient versus $IS$ measure for the artificial dataset.**

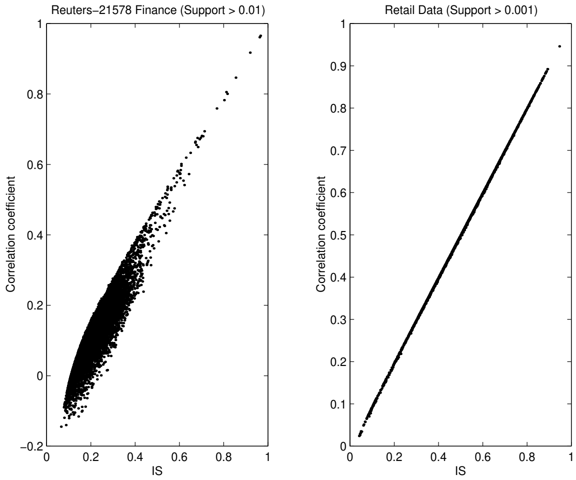


**Figure 5: Correlation coefficient versus $IS$ measure for Reuters-21578 (Finance) and retail dataset).**

newswire articles[2]. This dataset contains 2886 attributes and 2005 documents. The second dataset is obtained from a large retail corporation. This dataset has 14462 attributes and 58565 transactions. The relationship between $IS$ and $\phi$-coefficient for these datasets are shown in Figure 5.

### 3.3 Other Measures

We now describe three other interestingness measures that can be used for association patterns. They are the Gini index [4], Piatetsky-Shapiro's rule-interest [15] and conviction [7].

The Gini index for an association rule $A \longrightarrow B$ is given by

$$\begin{aligned} Gini = \quad & P(A)(P(B|A)^2 + P(\neg B|A)^2) + \\ & P(\neg A)(P(B|\neg A)^2 + P(\neg B|\neg A)^2) \\ & -P(B)^2 - P(\neg B)^2 \end{aligned} \qquad (12)$$

This value may range from 0 (when $A$ and $B$ are completely independent) to 0.5 (for perfect correlation).

The rule-interest function, which was introduced in [15], can be defined to be :

$$RI = P(A,B) - P(A)P(B) \qquad (13)$$

The range of this function is between -0.25 and 0.25.

Conviction was introduced in [7] as an asymmetric version of the interest factor.

$$\text{conviction} = \frac{P(A)P(\neg B)}{P(A,\neg B)} \qquad (14)$$

This measure is intuitively derived from interest factor in the following way. The rule $A \longrightarrow B$ is logically equivalent to $\neg(A \wedge \neg B)$. Since the interest factor between $A$ and $\neg B$ is $\frac{P(A,\neg B)}{P(A)P(\neg B)}$, equation 14 is obtained by inverting the ratio. This inversion is due to the outside negation symbol in the logical expression $\neg(A \wedge \neg B)$. The value of conviction ranges from 0 to $+\infty$.

## 4. RANKING OF ASSOCIATION PATTERNS

In this section, we will show how the various measures described previously can be used for ordering the association patterns according to their degree of interestingness. Instead of ordering every itemsets, a good starting point would be to rank only itemsets that fall into region II of Fig. 1. Firstly, we need to determine the maximal frequent itemset border and $\chi^2$ border using algorithms such as Apriori [3] and the Dependence Rules Algorithm[17]. The two borders can be used to remove all itemsets that are infrequent or independent. We would then compute the interest value for each remaining itemset according to an interestingness measure, $F$. If an analyst is only interested in itemsets, we can return the highly-ranked itemsets.

However, if an analyst is interested in rules rather than itemsets, one must define the corresponding objective measures for rules, $F'$. In many cases, the objective measures for itemsets may not be the same as that for association rules. Therefore, one must ensure that both $F$ and $F'$ are consistent with each other.

[2]available at http://www.research.att.com/~lewis.

We will now illustrate an example of ranking itemsets and rules using the interest factor, $I$. Consider a large k-itemset $\{A_1, A_2 \cdots A_k\}$. There are $2^{k-2}$ ways to partition the itemset into rules. [3] The interest factor for the large k-itemset is :

$$I(A_1, A_2, \cdots A_k) = \frac{P(A_1, A_2, \cdots A_k)}{P(A_1)P(A_2)\cdots P(A_k)} . \qquad (15)$$

Suppose we want to compute the interest factor for the rule $A_1A_2\cdots A_j \longrightarrow A_{j+1}A_{j+2}\cdots A_k$. We can rewrite the above equation into the following form :

$$\begin{aligned} & I(A_1, \cdots A_k) \\ = & \frac{P(A_1, A_2, \cdots A_j)\ P(A_{j+1}, A_{j+2}, \cdots A_k|A_1 \cdots A_j)}{P(A_1)P(A_2)\cdots P(A_k)} \\ = & \frac{P(A_1, A_2, \cdots A_j)}{P(A_1)P(A_2)\cdots P(A_j)} \ \frac{P(A_{j+1}, A_{j+2}, \cdots A_k|A_1A_2\cdots A_j)}{P(A_{j+1})P(A_{j+2})\cdots P(A_k)} \\ = & I(A_1, A_2, \cdots A_j)\ \frac{P(A_1, A_2, \cdots, A_n)}{P(A_{j+1})P(A_{j+2})\cdots P(A_k)P(A_1A_2\cdots A_j)} \\ & \times \frac{P(A_{j+1}, A_{j+2}, \cdots A_k)}{P(A_{j+1}, A_{j+2}, \cdots A_k)} \\ = & I(A_1, \cdots A_j)\ I(A_{j+1}, \cdots A_k) \\ & \times \frac{P(A_1, A_2, \cdots A_k)}{P(A_1, \cdots A_j)P(A_{j+1}, \cdots A_k)} . \end{aligned} \qquad (16)$$

The above equation allows us to define the interest factor for a rule in terms of the interest factor for the corresponding itemsets :

*Definition 1.* The interest factor for the rule

$$A_1A_2\cdots A_j \longrightarrow A_{j+1}A_{j+2}\cdots A_k$$

can be defined as :

$$\begin{aligned} & I(A_1\cdots A_j \longrightarrow A_{j+1}\cdots A_k) \\ \equiv & \frac{P(A_1, A_2, \cdots A_k)}{P(A_1, A_2, \cdots A_j)P(A_{j+1}, A_{j+2}\cdots A_k)} \\ = & \frac{I(A_1, A_2, \cdots A_k)}{I(A_1, A_2, \cdots A_j)\ I(A_{j+1}, A_{j+2}, \cdots A_k)} . \end{aligned} \qquad (17)$$

The above definition is useful because it allows us to compute the interest factor for a rule using only the interest factors of the itemsets. Furthermore, it says that the best rule for a given itemset is the one that maximizes the difference between $I(A_1, A_2, \cdots A_k)$ and the product $I(A_1, A_2, \cdots A_j)$ $I(A_{j+1}, A_{j+2}, \cdots A_k)$. This definition can also be used to define the interest part of the $IS$ measure for an association rule.

## 5. RESULTS

One way to compare the various measures presented in this paper is by determining their correlation with respect to the $\phi$-coefficient. Table 3 illustrates the correlation values computed using the artificial dataset, for various ranges of support values. For asymmetric measures such as confidence and conviction, we represent the confidence or conviction

[3]Here, due to the symmetry of the $I$ factor, we assume that the rules $A \longrightarrow B$ and $B \longrightarrow A$ have the same interest value.

Table 3: **Correlation between different interestingness measures and $\phi$-coefficient for various range of support values. These coefficients are computed for itempairs generated using the artificial dataset.**

| Support | Interest | $IS$ | laplace | conviction | confidence | $\lambda$ | entropy | RI | Gini index |
|---|---|---|---|---|---|---|---|---|---|
| [0, 1] | 0.7057 | 0.7981 | 0.7855 | 0.0511 | 0.7854 | -0.0027 | -0.0065 | 0.9811 | -0.0046 |
| [0.005, 1] | 0.7055 | 0.7979 | 0.7862 | 0.0510 | 0.7861 | 0.0136 | 0.0220 | 0.9814 | 0.0151 |
| [0.01, 1] | 0.7135 | 0.7974 | 0.7846 | 0.0510 | 0.7845 | 0.0353 | 0.0541 | 0.9818 | 0.0388 |
| [0.05, 1] | 0.7393 | 0.7915 | 0.7659 | 0.0534 | 0.7659 | 0.2101 | 0.2577 | 0.9840 | 0.2263 |
| [0.005, 0.7] | 0.7293 | 0.8627 | 0.8856 | 0.0477 | 0.8854 | 0.0555 | 0.1011 | 0.9911 | 0.0511 |
| [0.01, 0.7] | 0.7391 | 0.8650 | 0.8879 | 0.0476 | 0.8878 | 0.0738 | 0.1327 | 0.9912 | 0.0746 |
| [0.05, 0.7] | 0.7725 | 0.8760 | 0.8929 | 0.0476 | 0.8928 | 0.2506 | 0.3566 | 0.9921 | 0.2855 |
| [0.005, 0.5] | 0.7315 | 0.9318 | 0.9298 | 0.0483 | 0.9296 | 0.5280 | 0.5571 | 0.9800 | 0.4722 |
| [0.01, 0.5] | 0.7433 | 0.9342 | 0.9313 | 0.0480 | 0.9311 | 0.5401 | 0.5831 | 0.9798 | 0.4920 |
| [0.05, 0.5] | 0.7835 | 0.9505 | 0.9350 | 0.0458 | 0.9349 | 0.6970 | 0.7601 | 0.9777 | 0.6914 |
| [0.005, 0.3] | 0.7057 | 0.9806 | 0.9317 | 0.3199 | 0.9311 | 0.8644 | 0.9023 | 0.9492 | 0.8426 |
| [0.01, 0.3] | 0.7280 | 0.9820 | 0.9340 | 0.3193 | 0.9336 | 0.8696 | 0.9101 | 0.9469 | 0.8482 |
| [0.05, 0.3] | 0.7704 | 0.9871 | 0.9273 | 0.3076 | 0.9271 | 0.9147 | 0.9452 | 0.9316 | 0.8897 |

value of an itempair by the maximum value for all the rules generated from the itempair.

Notice that the correlation between RI and $\phi$ consistently stays above 0.9 for all the support regions considered in Table 3. However, for the last three rows, $IS$ seems to be the best choice, which is not surprising considering it is derived from the correlation coefficient itself. On the other hand, conviction works poorly even for the low support region. This is because it has a very wide range of values (from 0 to $\infty$). Other measures such as the $\lambda$-coefficient and Gini index have very low correlation with $\phi$ when no support thresholds are imposed. This is because both measures are symmetric about zero (i.e. their values are non-negative). However, as the support region becomes smaller, the symmetry will be broken and the correlation values become larger (Fig. 6).

Finally, note that many of the interestingness measures are highly correlated with the $\phi$ coefficient for the last three rows of Table 3. This include the laplace function, maximum confidence, $\lambda$, entropy and the Gini index.

## 6. CONCLUSIONS

The following conclusions can be made :

- Support is a good measure because it represents how statistically significant a pattern is. Support-based pruning is effective because it allows us to prune mostly uncorrelated or negatively correlated patterns.
- $\chi^2$ is appropriate to test whether there is sufficient evidence to show that items in a pattern are independent of each other. However, it does not quantify the strength of correlation among the items.
- Many of the measures (such as IS, laplace, maximum confidence, RI) have similar behavior in the region of medium support values (which typically occurs in many practical datasets). They provide similar information regarding the dependencies between items as correlation coefficient.

The above conclusions suggest that we can use any of these interestingness measures to rank patterns that belong to region II of Fig. 1. A good interestingness measure should be highly correlated with statistical correlation and takes into account the support of the pattern.

## 7. REFERENCES


[1] R. Agrawal, T. Imielinski, and A. Swami. Database mining: a performance perspective. *IEEE Transactions on Knowledge and Data Engineering*, 5:914–925, 1993.

[2] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. ACM SIGMOD Intl. Conf. Management of Data*, 1993.

[3] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, 1994.

[4] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Chapman & Hall, 1984.

[5] Tom Brijs, Gilbert Swinnen, Koen Vanhoof, and Geert Wets. "using association rules for product assortment decisions : A case study. In *Proc. of the Fifth Int'l Conference on Knowledge Discovery and Data Mining*, 1999.

[6] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proc. ACM SIGMOD Intl. Conf. Management of Data*, 1997.

[7] S. Brin, R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proc. of 1997 ACM-SIGMOD Int. Conf. on Management of Data*, 1997.

[8] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proc. of 1997 ACM-SIGMOD Int. Conf. on Management of Data*, 1997.

[9] Robert Cooley Chris Clifton. Topcat: Data mining for topic identification in a text corpus. In *Proceedings of the 3rd European Conference of Principles and Practice of Knowledge Discovery in Databases*, 1999.

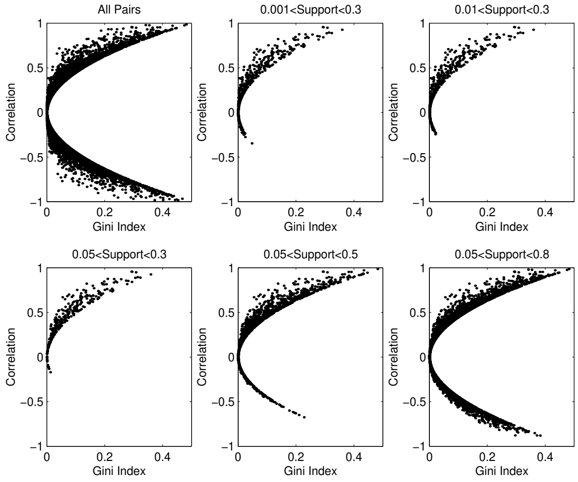


**Figure 6: Correlation coefficient versus Gini index for the artificial dataset.**


[10] Peter Clark and Robin Boswell. Rule induction with cn2 : Some recent improvements. In *Proceedings of the European Working Session on Learning EWSL-91*, 1991.

[11] L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications, ii: Further discussion and references. *Journal of the American Statistical Association*, 1959.

[12] B. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *Proc. of the Fifth Int'l Conference on Knowledge Discovery and Data Mining*, 1999.

[13] R. Ng, L. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained association rules. In *Proc. of 1998 ACM-SIGMOD Int. Conf. on Management of Data*, 1998.

[14] J.S. Park, M.S. Chen, and P.S. Yu. An effective hash-based algorithm for mining association rules. *SIGMOD Record*, 25(2):175–186, 1995.

[15] Gregory Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In Gregory Piatetsky-Shapiro and William Frawley, editors, *Knowledge Discovery in Databases*, pages 2299–248. MIT Press, Cambridge, MA, 1991.

[16] H.T. Reynolds. *The Analysis of Cross-Classifications*. The Free Press, New York, 1977.

[17] C. Silverstein, S. Brin, and R. Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2(1):39–68, 1998.

[18] R. Srikant and R. Vu, Q.and Agrawal. Mining association rules with item constraints. In *Proc. of the Third Int'l Conference on Knowledge Discovery and Data Mining*, 1997.

[19] H Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen, and H. Mannila. Pruning and grouping discovered association rules. In *ECML-95 Workshop on Statistics, Machine Learning and Knowledge Discovery in Databases*, 1995.

[20] R.L. Winkler and W.L. Hays. *Statistics: Probability, Inference and Decision*. Holt, Rinehart and Winston, second edition, 1975.