An Empirical Study of Qualities of Association Rules from a Statistical View Point

Maryann Dorn*, Wen-Chi Hou*, Dunren Che*, and Zhewei Jiang*

Abstract: Minimum support and confidence have been used as criteria for generating association rules in all association rule mining algorithms. These criteria have their natural appeals, such as simplicity; few researchers have suspected the quality of generated rules. In this paper, we examine the rules from a more rigorous point of view by conducting statistical tests. Specifically, we use contingency tables and chi-square test to analyze the data. Experimental results show that one third of the association rules derived based on the support and confidence criteria are not significant, that is, the antecedent and consequent of the rules are not correlated. It indicates that minimum support and minimum confidence do not provide adequate discovery of meaningful associations. The chi-square test can be considered as an enhancement or an alternative solution.

Keywords: *Data Mining, Association Rule Mining, Rule Evaluation, Chi-square Test*

1. Introduction

Mining market basket data [1, 2, 3, 7] has received a great deal of attention in the recent past, partly due to its utility and partly due to the research challenges it presents. Market basket data typically consists of store items purchased on a per-transaction basis, but it may also consist of items bought by a customer over a period of time. The goal is to discover buying patterns, such as two or more items that are often bought together. Such finding could aid in business decision making. Association rules reflect a fundamental class of patterns that exist in the data. Consequently, mining association rules in market basket data has become an important problem in data mining.

In general, the task of mining all rules (or association rules) can be accomplished in two steps:

(1) Find all *itemsets* that are above a given minimum support ratio. The support ratio for an *itemset* is the percentage of the number of transactions that contain the *itemset* against the total number of transactions. *Itemsets* satisfying the minimum support (*minsup*) are called large (or frequent) *itemsets*, and all others small *itemsets*. This step is responsible for most of the computation time, and has been the focus of considerable work in developing fast algorithms [1, 2, 3, 6, 7, 16, 19, 21]. Agrawal & Srikant [2, 3] have provided the initial foundation for this research problem.

(2) Use the large *itemsets* to generate the association rules. For example, both {A, B, C, D} and {A, B} are large *itemsets*. The association rule, AB => CD, is derived if at least c% of the transactions that contain AB also contain

CD, where c% is a pre-specified constant called minimum confidence (*minconf*).

There have been many algorithms developed to find association rules [1, 2, 3, 4, 5, 6, 11, 13, 17, 18, 21]. While association rule mining algorithms may be different in their efficiency, they all use *minsup* and *minconf* as the criteria to determine the validity of the rules. Due to their simplicity and natural appeals, few researchers have suspected the sufficiency of these criteria. In this research, we shall examine the rules derived based on these criteria in a more rigorous way by conducting statistical tests. The experimental results have shown that a surprising 30% of the rules satisfying the *minsup* and *minconf* are indeed insignificant statistically.

The rest of the paper is organized as follows. In Section 2, we present some background on statistical testing. Section 3 discusses the experimental setup and the results. Section 4 is the conclusions.

2. Background of Chi-Square Test for Independence

Consider the market basket example from [6]. The focus is on the purchase of tea and coffee. Assume the *minsup* is 5% and *minconf* is 50%. In Table 1, rows *tea* and *no_tea* corresponds to baskets (or transactions) that do and do not, respectively, contain *tea*; and similarly columns *coffee* and *no_coffee* corresponds to baskets that do and do not contain *coffee*. The numbers in the table represent percentage of baskets. A potential association rule '*tea=> coffee*' reads: "When people buy tea, they are also likely to buy coffee". The support for this rule is 20%, which is fairly high. The confidence, defined by the conditional probability that a customer buys coffee, given that he/she also buys tea, i.e., p[tea and coffee] / p[tea], is 20 / 25 = 0.8 or 80%, which is also high. Since the support and confidence satisfy the

pre-specified *minsup* (5%) and *minconf* (50%), the rule is accepted as valid in the support-confidence framework.

Table 1. A 2 by 2 Contingency Table.

	coffee	no_coffee	$\sum$ row
tea	20	5	25
no_tea	70	5	75
$\sum$ col	90	10	100

The chi-square (χ^2) test is a non-parametric statistical method that can be used to test independence among attributes. It is reliable under a fairly permissive set of assumptions. This approximation breaks down when the expected values are small. As a rule of thumb, statistics texts [9, 15] recommend the use of chi-square test only if (1) all cells in the contingency table have expected values greater than 1, and (2) at least 80% of the cells in the contingency table have expected values greater than 5. With small expected cell values, the usual alternative is to combine or collapse cells [9, 10, 15].

The χ^2 statistics is approximately distributed with a chi-square distribution, which is often tabulated in statistics texts. The χ^2 value is computed from a contingency table, which records the counts or frequencies of different combinations of attribute values. The χ^2 statistics has a parameter called degrees of freedom associated with it. In Table 1, we show a 2 by 2 contingency table.

Given an r×c contingency table, the χ^2 is computed as follows:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where $\sum$ is taken over all cells of the table; O_{ij} is the observed count for cell *ij*, and E_{ij} is the expected count for cell *ij*, which is calculated as $E_{ij} = n_i n_j / n$, where n_i, n_j are the total counts of row *i* and column *j*, respectively, and *n* is the total count of all the cells. The table has a (r-1)×(c-1) degree of freedom.

When Table 1 is tested by chi-square statistics, the result, $\chi^2 = 3.703703$ with degree of freedom = 1, is non-significant at 95% confidence level; it indicates that tea and coffee are independent variables. In other words, tea is not a determining factor whether people will buy coffee or not. Obviously, there is some discrepancy between the conclusions drawn from the two methods.

Potential advantages of the χ^2 statistics over the commonly used support-confidence framework [6] are:

(1) The use of the chi-square significance test for independence is more solidly founded in statistical theory. In particular, there is no need to choose ad-hoc values of support and confidence.

(2) The chi-square statistic simultaneously and uniformly takes into account all possible combinations of the presence and absence of the various attributes being examined as a group.

(3) The chi-square test at a given significance level is upward closed. In other words, if an i-itemset is correlated, all its supersets are also correlated.

3. Experiments

In this section, we present the experimental design and analyze its results.

3.1 Experimental Design

Four synthetic data sets are generated using the IBM/Quest data generator [3], which has been widely used for evaluating association rule mining algorithms. The data sets generated are then fed into CBA/DBII data mining system [12] to generate association rules. Finally, contingency tables are constructed using Microsoft Excel, and the chi-square statistical tests are conducted on the association rules generated.

The synthetic transactions generated are supposed to mimic the transactions in the retail environment. The model of the "real" world is that people tend to buy sets of items together. A transaction may contain more than one large *itemset*. For example, a customer might place an order for a dress and jacket when ordering sheets and pillowcases, in which the dress and jacket form one large *itemset* and sheets and pillowcases together form another large *itemset*. Transaction sizes are typically clustered around a mean with a few transactions having many items. Typical sizes of large *itemsets* are also clustered around a mean, with a few *itemsets* having a large number of items.

Some of the options of the data generator are as follows:

-ntrans
-tlen   Average items per transaction (default: 10)
-nitems
-npats   number of patterns (default: 10000)
-patlen average length of maximal pattern (default = 4)
-corr   correlation between patterns (default 0.25)
-conf   average confidence in a rule (default = 0.75)

The transaction size is picked from a Poisson distribution [3] with mean μ equal to -tlen. Note that if each item is chosen with the same probability P and there are N items, the expected number of items in a transaction is given by a binomial distribution with parameters N and P, and is approximated by a *Poisson* distribution with mean NP.

In order to model the phenomenon that large *itemsets* often have common items, some fraction of the items in the subsequent *itemsets* are chosen from the previous *itemset* generated. It uses an exponentially distributed random variable [3] with mean equal to the correlation level *(-corr)* to decide this fraction for each *itemset*. The remaining items are picked at random. In the datasets applied in this study, the correlation level was set to 0.25. Agrawal & Srikant [3] ran some experiments with the correlation level set to 0.5 and 0.75 but did not find much difference in the nature of the performance results.

Each *itemset* in the database has a weight associated with it, which corresponds to the probability that this

itemset will be picked. This weight is picked from an exponential distribution with unit mean, and is then normalized so that the sum of the weights for all the *itemsets* is 1.

Four synthetic datasets are generated with slightly different options for comparison, such as, number of transactions, number of items, and average confidence for rules (see Table 2). The other common factors are 'average transaction length (5)', 'number of patterns (100)', 'average length of pattern (4)', and 'correlation between consecutive patterns (0.25)'.

Table 2: Four Datasets

Data sets	Num of transactions	Num of items	Avg. Conf. level
1	2217	20	0.5
2	5188	20	0.5
3	5029	40	0.5
4	5096	40	0.75

Once the data sets are generated, the CBA system [12, 14] is used to generate association rules that satisfy the given *minsup* and *minconf*.

3.2 Experimental Results and Chi-Square Test

First, we discuss the effects of *minsup* and *minconf* on the numbers of large *itemsets* and rules produced. Then, we will examine the quality of the rules using the chi-square test.

3.2.1 Effects of Minsup and Minconf

The value of minimum support (*minsup*) controlls the number of large *itemsets*. Table 3 shows the drastic reduction of the number of large *itemsets* when *minsup* increases from 1% to 5%. For example, in the 2217 transaction dataset, the number of large *itemsets* drops from 3401 to 408.

The size of dataset can also affect the level of *minsup* used significantly. The larger the data size, the smaller the *minsup* is needed to provide a controllable size of large *itemsets*. For example, the datasets with over 5000 transactions generate fewer large *itemsets* than the dataset with 2217 transactions for the same *minsup* (see Table 3).

Table 3. Number of Large Itemsets at Various Minsup Levels

Data set	1%	2%	5%	10%	15%	20%
1 (2217)	3401	1488	408	131	65	36
2 (5188)	1836	808	246	84	37	24
3 (5029)	1513	521	118	40	17	11
4 (5096)	1600	531	118	40	17	11

The relationship among *minsup*, *minconf*, and the number of the association rules is shown in Tables 4, 5, 6, and 7 for each dataset respectively. The confidence level also affects the number of association rules generated significantly. The higher the level of *minconf* becomes, the smaller the number of association rules sustains. The determination of *minconf* cutoff point is a major problem. When the data size is large, a lower *minsup* is preferred. If the *minsup* is set too high, combined with a *minconf* at 75% or 90%, often it generates no association rule. Table 5 shows that at *minsup*=1% and *minconf*=50%, it only generates 1487 rules, and with *minconf*=75% it generates 217 rules. But when the *minsup* moves up to 2%, with *minconf*=75%, it only generates 43 rules.

Table 4. Number of Transactions: 2217, Average Confidence: 0.5

minsup	1%			5%			10%			15%		
minconf	50%	75%	90%	50%	75%	90%	50%	75%	90%	50%	75%	90%
Itemsets	3401	3401	3401	408	408	408	131	131	131	65	65	65
Rules	3169	1261	89	358	38	0	112	3	0	50	0	0

Table 5. Number of Transactions: 5188, Average Confidence: 0.5

minsup	1%			5%			10%			15%		
minconf	50%	75%	90%	50%	75%	90%	50%	75%	90%	50%	75%	90%
Itemsets	1836	1836	1836	246	246	246	84	84	84	37	37	37
Rules	1487	217	9	176	5	0	57	0	0	19	0	0

Table 6. Number of Transactions: 5029, Average Confidence: 0.5

minsup	1%			5%			10%			15%		
minconf	50%	75%	90%	50%	75%	90%	50%	75%	90%	50%	75%	90%
Itemsets	1513	1513	1513	118	118	118	40	40	40	17	17	17
Rules	576	247	30	21	0	0	6	0	0	3	0	0

Table 7. Number of Transactions: 5096, Average Confidence: 0.75

minsup	1%			5%			10%			15%		
minconf	50%	75%	90%	50%	75%	90%	50%	75%	90%	50%	75%	90%
Itemsets	1600	1600	1600	118	118	118	40	40	40	17	17	17
Rules	677	204	64	24	0	0	8	0	0	3	0	0

3.2.2 Chi-Square Tests for Independence

The results from the chi-square tests for independence constitute the major findings of this study. Note that all chi-square tests conducted in this study adopted the commonly used significance level, $p \leq 0.05$, as a cutoff point. The critical value is 3.841 for 1 degree of freedom.

In the following, we use the dataset with 2217 transactions as an example. Each transaction has three attributes: Transaction number, Customer ID, and item. To facilitate interpretation, 20 items are randomly assigned different labels.

The first experiment was conducted under *minsup*=10% and *minconf*=50%. CBA produced 131 large itemsets and 112 association rules as shown in Table 4 under the 10% (*minsup*) and 50% (*minconf*) column. All 112 rules generated were then tallied into 2 by 2 contingency tables and tested by the chi-square test for independence. The results show that 37 out of 112 rules are not significant, in other words, 37 rules show that there is no relationship between their antecedents and the consequents.

With the same dataset, when the *minsup* was reduced to 5% and the *minconf* was kept the same level at 50%, that is, the restriction was loosened, the CBA program produced 408 large itemsets and 385 association rules as shown in Table 4 in the 5% and 50% column. Out of these 385 rules, 134 rules are not significant. The previous 37 rules are part of these 134 rules. It again shows that no relation between the antecedents and consequents in one third of the association rules derived by the support-confidence framework. The results demonstrate a striking discrepancy between the support-confidence framework and the chi-square statistical tests.

3.3 Analysis

We use examples to see why some of the rules generated according to *minsup* and *minconf* may not be valid statistically.

3.3.1 Sufficiency of Minsup and Minconf

The following two examples illustrate one interesting phenomenon. The support and the confidence ratios for Rule 175 (Example 1) and 169 (Example 2) are exactly the same. With the same cut-off *minsup* and *minconf*, the two rules are both valid. However, they have the opposite fate in statistical tests. The χ^2 value for Rule 175, as shown in Table 8, is not significant ($p > 0.05$). It indicates that 'buying apple and ketchup' is independent from 'buying

Example 1. Non-Significant Relationship
Table 8. Chi-square Test vs. Minsup/Minconf

Rule 175: AP/KP =>PA					
	R	NO_R	T_ROW	SUP_L	CONF
L	111	74	185	8.34%	60.00%
NO_L	1134	898	2032		
T_COL	1245	972	2217		
SUP_R	56.16%			5.01%	
$p \leq 0.271182286$					

Example 2. Significant Relationship
Table 9. Chi-square Test vs. Minsup/Minconf

Rule 169: AP/KP => OV					
	R	NO_R	T_ROW	SUP_L	CONF
L	111	74	185	8.34%	60.00%
NO_L	1037	995	2032		
T_COL	1148	1069	2217		
SUP_R	51.78%			5.01%	
$p \leq 0.019456897$					

pamper'. In Example 2, the χ^2 value for Rule 169 is significant ($p < 0.05$). It indicates that there is a relationship between 'buying an oven' and 'buying apple and ketchup'. It implies that merely depending on *minsup* and *minconf* as a constraint factor may not provide a reliable basis for judging the association.

3.3.2 Uni-Directional vs. Bi-Directional

The association rule in the Apriori algorithm is uni-directional. A rule X => Y does not necessarily imply the rule Y => X. In addition, it cannot detect negative implication, such as, buying product X and not buying product Y. However, the chi-square statistic simultaneously and uniformly takes into account all possible combinations of the presence and absence of the various attributes. We shall illustrate these in the following examples.

Example 3. Bi-directional Relationship
Table 10. Chi-square Test on Bi-directional Relationship

Rule 1: T => S					
	S	NO_S	T_ROW	SUP	CONF
T	36	19	55	29.10%	65.45%
NO_T	112	22	134		
T_COL	148	41	189		
	78.31%			19.05%	
$\chi^2 = 7.5433048$		P $\leq$ 0.01			

(a). Rule: T=> S

Rule 2: S => T					
	T	NO_T	T_ROW	SUP	CONF
S	36	112	148	78.31	24.32%
NO_S	19	134	41		
T_COL	55	134	189		
	29.10%			19.05	
$\chi^2 = 7.5433048$		P $\leq$ 0.01			

(b). Rule: S=> T

Example 3 shows that at *minconf* =50%, the confidence of Rule 1: T => S is high (65.45%), so Rule 1 is accepted as an association rule. But the confidence of Rule 2: S=> T is too low (24.32%), so Rule 2 is pruned off. However, the χ^2 values of both Rule 1 and Rule 2 are the same and their P values are all less than 0.05, which implies that T and S are correlated and S=>T should have been accepted as a rule.

It is also useful that the chi-square test can detect negative implication. Example 4 illustrates this case. The χ^2 values for both L => R and L => (NO_R) are the same. Since $p > 0.05$, it indicates that L and R are two independent

Example 4. Negative Relationship.
Table 11. Chi-square Test on Negative Relationship

Rule 3: L => R							RULE 3: (Expected)		
	R	NO_R	T_ROW	SUP	CONF			R	NO_R
L	29	26	55	29.10%	52.73%		L	30.85	24.15
NO_L	77	57	134				NO_L	75.15	58.85
T_COL	106	83	189						
	56.08%				15.34%				
p ≤ 0.55128125							p ≤ 0.55128125		

(a) Rule 3: L => R

Rule 4: L => NO_R							RULE 4: (Expected)		
	NO_R	R	T_ROW	SUP	CONF			NO_R	R
L	26	29	55	29.10%	47.27%		L	24.15	30.85
NO_L	57	77	134				NO_L	58.85	75.15
T_COL	83	106	189						
	43.92%				13.76%				
p ≤ 0.55128125							p ≤ 0.55128125		

(b) Rule 4: L => NO_R

attributes. However, at *minconf* =50%, the confidence of the Rule 3: L => R is high (52.73%), so it is accepted as an association rule; but the confidence of Rule 4: L=> NO_R is low (47.27%), so Rule 4 is pruned off.

3.4 Enhancing the Support-confidence Framework

The support-confidence framework can be enhanced by incorporating the chi-squared test. In fact, the required statistics for conducting chi-squared tests can be obtained with little effort as follows. Let |T| be the total number of transactions in the database. Let X be an *itemset* and $|T_X|$ (or the support of X) the number of transactions containing X. Let $X \cup Y$ be an *itemset* that contains all the items in X or Y. Assume X, Y, and $X \cup Y$ are identified as frequent *itemsets* in the process of deriving frequent itemsets. Then, the required statistics for X_but_not_Y can be computed, based on the set theory, as $|T_X|$ -$|T_X \cup Y|$, Y_but_not_X as $|T_Y|$-$|T_X \cup Y|$, and not_X_and_not_Y as $|T|$ -$|T_X|$-$|T_Y|$+$|T_X \cup Y|$. Therefore, given a minimum support requirement, all rules derived by the support-confidence framework can be tested using the chi-square statistics on the fly without much effort. With the incorporation of the chi-squared tests, the derived rules should be more reliable statistically.

4. Conclusions

We present in this paper an empirical study on the validity of association rules derived based on the *minsup* and *minconf* criteria. The results of this study strongly suggest that *minsup* and *minconf* do not provide adequate discovery of associations. From the results of chi-square tests, we conclude as follows.

(1) *Minsup* and *minconf* can significantly cut down the number of rules generated but does not necessarily only cut off those insignificant rules.

(2) Association rules derived are not very reliable in making inference such as "people who buy product A tend to buy product B." Approximately, one-third of such association rules, discovered with a *minconf* at 50% and *minsup* at either 5% or 10%, fail to show that there exists statistical relationship between the antecedents and the consequents.

(3) The relationship between *minsup* and *minconf* was not decisive. High level of *minconf* (70% above) in general implies significant relationship with either high or low *minsup*. Mid range of *minconf* mixed with either mid or high level of *minsup* could produce either significant or not-significant outcomes.

The chi-squared test has demonstrated itself as a more prudent way to discover correlations. The test can also be easily incorporated into the support-confidence framework (as discussed in Section 3.4). Combining the efficient algorithms for deriving frequent itemsets of the support-confidence framework with the theoretically sound chi-squared test, building an efficient and reliable data mining system is achievable.

References

[1] Agrawal, R., Imielinski, T., and Swami, A. "Mining Association Rules Between Sets of Items in Large Databases," In Proc. of the ACM-SIGMOD Conf. on Management of Data, Washington, D. C., 1993, pp. 207-216.

[2] Agrawal R., Srikant, R. "Fast algorithms for Mining Association Rules," In Proc. of the 20th VLDB

Conference, Santiago, Chile, 1994, pp. 487-499.

[3] Agrawal, R. and Srikant, R. "Fast Algorithms for Mining Association Rules," IBM Research Report RJ9839, June 1994. IBM Almaden Research Center, San Jose, CA.

[4] Bayardo, R. J. and Agrawal, R. "Mining the Most Interesting Rules," In Proc. of the Fifth ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 1999, pp.145-154.

[5] Bayardo, R., Agrawal, R, and Gunopulos, D. "Constraint-Based Rule Mining in Large, Dense Databases," In Proc. of the 15th Int'l Conf. on Data Engineering, 188-197, 1999.

[6] Brin, S. Motwani, R. and Silverstein, R. "Beyond Market Basket: Generalizing Association Rules to Correlations." SIGMOD-97, 1997, 265-276.

[7] Brin, S., Motwani, R., Ullman, J., and Tsur, S. "Dynamic Itemset Counting and Implication Rules for Market Basket Data." In Proc. of the 1997 ACM-SIGMOD Int'l Conf. on the Management of Data, 1997, 255-264.

[8] Ganti, V., Gebrke, and Ramakrishnan, R. "Mining Very Large Databases," Computer, Vol. 32, No. 8, Aug. 1999, pp. 38-45.

[9] Glass, G. V. and Hopkins, K. D. Statistical Methods in Education and Psychology. (2nd ed.) Prentice Hall, New Jersey, 1984.

[10] Gokhale, D. V. and Kullback, S. The Information in Contingency Tables. Marcel Dekker Inc., New York, 1978.

[11] Han, J. and Fu, Y. "Discovery of multiple-level association rules from large databases." VLDB-95.

[12] Liu B., Hsu W., and Ma Y. "Pruning and Summarizing the Discovered Associations, " in Proc. of the ACM SIGKDD Int'l Conference on Knowledge Discovery & Data Mining, San Diego, CA, 1999.

[13] Liu B., Hsu W., and Ma Y. "Mining Association Rules with Multiple Minimum Supports" in Proc. of the ACM SIGKDD Int'l Conference on Knowledge Discovery & Data Mining, 1999.

[14] Liu B., Hsu W., Wang K., and Chen S. "Mining Interesting Knowledge Using DM-II" in Proc. of the ACM SIGKDD Int'l Conference on Knowledge Discovery & Data Mining, 1999.

[15] Mason, R. D., Lind, D. A., and Marchal, W. G. STATISTICS: An Introduction, 5th ed. Duxbury Press, 1998.

[16] Park, J. S.; Chen, M.-S.; and Yu, P. S. An Effective Hash Based Algorithm for Mining Association Rules. In Proc. of SIGMOD Conf. on the Management of Data, 1995, pp 175-186.

[17] Srikant, R. and Agrawal, R. "Mining Generalized Association Rules," In Proc. of the 21st Int'l Conf. on VLDB, 1995, pp. 407-419.

[18] Srikant, R. and Agrawal, R. Mining Generalized Association Rules. IBM Research Report RJ9963, June 1995. IBM Almaden Research Center, San Jose, CA.

[19] Srikant, R., Vu, Q., and Agrawal, R. "Mining Association Rules with Item Constraints," In Proc. of the Third Int'l Conf. on Knowledge Discovery in Databases and Data Mining, 1997, pp. 67-73.

[20] Toivonen H. "Sampling Large Databases for Association Rules," In Proc. of the 22th VLDB Conference, Mumbai, India, 1996, pp. 134-144.

[21] Zaki, M. J.; Parthasarathy, S.; Ogihara, M.; and Li, W. New Algorithms for Fast Discovery of Association Rules. In Proc. of the Third Int'l Conf. on Knowledge Discovery in Databases and Data Mining, 1997, pp. 283-286.

Maryann Dorn
She received a M.S degree in Computer Science from Southern Illinois University in 2000.
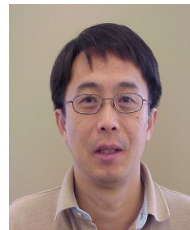
Wen-Chi Hou
He received a Ph.D. degree in Computer Sci. & Eng. from Case Western Reserve Univ., Cleveland Ohio, in 1989. He is an associated professor at Southern Illinois University Carbondale, USA. His interests are in database & data mining.

Dunren Che
He received his Ph.D. in Computer Science from BUAA China in 1994. He is an assistant Professor at Southern Illinois University Carbondale, U.S.A. His main interests are in database and data mining.

Zhewei Jiang
She is currently a Ph.D. student in the Department of Computer Science at Southern Illinois University, Carbondale Illinois, USA. Her research interests include database and data mining.