The Fuzzy Search for Association Rules with Interestingness Measure

Phaichayon Kongchai, Nittaya Kerdprasop, and Kittisak Kerdprasop

***Abstract*—Association rule are important to retailers as a source of knowledge to manage shelf, to plan an effective promotion, and so on. However, when we are mining with association rule discovery technique, we normally obtain a large number of rules. To select only good rule is difficult. Therefore, in this paper we propose the fuzzy search technique to discover interesting association rule. The comparative result of fuzzy versus non-fuzzy searches are presented in the experiment section. We found that fuzzy search is more flexible than the non-fuzzy one in finding highly constrained rules.**

***Index Terms*—Fuzzy set, fuzzy search, membership function, association rule mining.**

I. Introduction

Association rule mining is a method to discover the patterns of information, such as the pattern to reveal that there are many people coming in the supermarket to buy some specific set of products. Therefore, the owner would like to know the buying patterns of customers. The owner should perform by 2 steps, first step, to records the purchase of individual customers in the tables. Second step, when get enough information then bring it to association rule mining and then the results are association rules. This method is called “Market Basket Analysis” and this association rules are usually used in business [1]. Therefore, to select the appropriate association rules to apply, it is necessarily very much and the researcher [2] proposes an algorithm to search with many constraints and can be reduced the search space.

But the most researchers continue to straightforward search association rules with normal constraint, For example if the user required support value to be equal 1.0 and items in the ‘then’ must be X (The variable X means items that the user wants.) [3]-[6] this searching technique is less efficient than fuzzy searching technique because the results must be support value as 1.0 only (Which makes it does not received the close results, such as support 0.99 but the fuzzy searching technique may be received the results with that support 0.99.). The Fuzzy set is very popular in a variety of major because it can indicate the level of what is uncertain. There are many researchers used it during processing, such as a data support system for sales promotion analysis using fuzzy query [7] this work used technical fuzzy and used probability to search sales information by SQL language. Which the original searching was not able to searched some information but fuzzy searching and weight with probability they can do it. There are many tasks related to using SQL and fuzzy in that searching [8], [9].

This research proposed the method to search association rules by applying fuzzy set and search them from the measure performance of association rule (Support and Confident). We also proposed technique to select and to rank the association rules with the scores, therefore the results are very satisfactory in some case.

II. Related Theories and Studies

A. Basic of Association Rule

Association rule is a data mining to discover the patterns or relationships of items from the large database [10]. Which is can be using association rules to predict something happen in the future.

Example 1

Beer, Coke => Diaper
or
If Beer, Coke then Diaper

It means if people who buy beer and coke then buy diaper together. Therefore the association rule is important to business or something that need to find relationships. To create the association rules that have two steps.

Step 1, Find all frequent itemsets meaning itemsets whose greater than or equal minimum support, and then we can find the support with equation (1).

Support(A) = all transaction that contain A/all transaction (1)

Step 2, Generate association rules with the frequent itemsets whose greater than or equal 2-itemsets and the association rules must be greater than or equal minimum confidence. We can find the confidence of the association rules by an equation (2).

Confidence(A=>B) = support(A and B)/support(A) (2)

B. Fuzzy Set

Fuzzy sets are two words that come from the word "fuzzy" that mean something is not clear, such as, The feeling of people are differently to discriminate, such as someone open air at 15 degrees, that is cool, but other people said that be very cold, and another word is "set" in this case mean a set of

Manuscript received December 13, 2013; revised March 1, 2014. This work was supported in part by grant from Suranaree University of Technology through the funding of Data Engineering Research Unit.

The authors are with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: Zaguraba_ii@hotmail.com).

DOI: 10.7763/IJCTE.2014.V6.915

mathematical sets that are composed a different member of the set, for instance, that consists of people, animals and objects. Therefore the term of "fuzzy sets" mean the sets that are composed a fuzzy member. This uncertainty, we cannot say whether it is true or not true, but we can tell level of the fact by membership functions.

In detail of how to compute the membership functions we explained in the next section. However, we can computed the relationship of fuzzy sets by 3 fuzzy set operations [11].

1) Fuzzy Complements may be called a complement of any set, which is a set has a relationship with another set, and then we can compute this relationship by equation as follows (3).

$$\mu_A - (X) = 1 - \mu_A(X) \tag{3}$$

2) Fuzzy Intersections is to extract the duplicate memberships of two sets or more, and then we can compute this relationship by equation as follows (4).

$$\mu_{A\cap B}(X) = \min[\mu_A(X), \mu_B(X)] \tag{4}$$

3) Fuzzy Unions are to include the fuzzy members of sets with the relationship. By equation as follows (5).

$$\mu_{A\cup B}(X) = \max[\mu_A(X), \mu_B(X)] \tag{5}$$

In this paper we used two operators to be computed the relationship of fuzzy sets are fuzzy intersections and fuzzy unions.

III. The Fuzzy Searching Association Rules Technique

The methodology of this research to search the association rules with fuzzy set technique is composed of three steps: (Fig. 1) first step user is defining user-constraint to search association rules from measure support and confidence, The second step computes a membership value of the association rules from membership function, and the last step the results were to calculate the score for select appropriate association rules and ranking, details of all the methods to describe as follows.

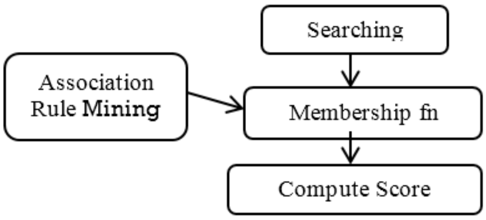

Fig. 1. The process of fuzzy search for association rules.

A. Searching

Association rule discovery based on user-constraint, the user can be defined the support and confidence value of the association rules by using the word "approximately", "approximately more than", "approximately less than", "and" and "or".

Example 2
User require association rules with the support *approximately* 0.7 and
The confidence approximately 0.8

Example 3
User require association rules with the support *approximately less than* 0.7 or
The confidence approximately 0.8

Example 4
User require association rules with the support *approximately* 0.7 and
The confidence approximately more than 0.8

From examples, you can be seen that the words (the words use Italics) used are different from normal search (Example, equal, less than, more than) because the user can define the conditions in the word form of approximately (Notice: the numbers are defined as 0.7 and 0.8 because we make it easier to explain, but you can change the value that you want.).

B. Association Rule

This research used association rules are the input to search the approximately rules. By association rules will be formatted as if – then. And the "if" and "then" contain information are call item, such as if A and B then C, if A and B then C and D. In addition, each rule has relation values that indicate the quality of association rules. Typically, most people used the support and confidence as the criterion to selected association rules, the support measure will indicate a number of transactions or the number of rows that support association rules, and the confidence measure can indicate the validity of the rules, which these measures will be important in determining selection association rules.

TABLE I: The Example of Association Rules.

NO.	Rules	Support	Confidence
1	G then F	0.1	0.4
2	B then C	0.2	0.44
3	F then E	0.3	0.45
4	A or B then C	0.48	0.56
5	A and B then E	0.5	0.58
6	G and B then E	0.6	0.59
7	F and T and B then C	0.79	0.62
8	A and B and G then F	0.83	0.75
9	G and R then F and B	0.9	0.76
10	A and F then C and B	1	0.8

C. Membership Functions

The Membership functions are intended to indicate the degree of the fuzzy sets. In this research we choose three functions are Triangular membership function (Fig. 2 a), R membership function (Fig. 2 b) and L membership function (Fig. 2 c). Which we will choose the triangulation membership function for conditions with the word "approximately" because the center point of the function

(Fig. 2 a, point b), with a maximum membership value is 1 which corresponds a user-defined the word about it, the R membership function is selected when the condition has the word “approximately less than” because of the value of most preferred users, they must be less than or equal to the value of user-defined then they have the membership value is 1, and the values which greater than the value of user-defined that will be had the membership value is reducing, and the L membership function be selected when the condition has the word “approximately more than” because of the value of most preferred users, they must be more than or equal to the number of user-defined then they have the membership value is 1, and the values whose less than the value of user-defined that will be had the membership value is reducing. From the Example 4 the user requires the association rules with the support approximately 0.8 and the confidence rather than 0.7. In the process of membership can be achieved by the following.

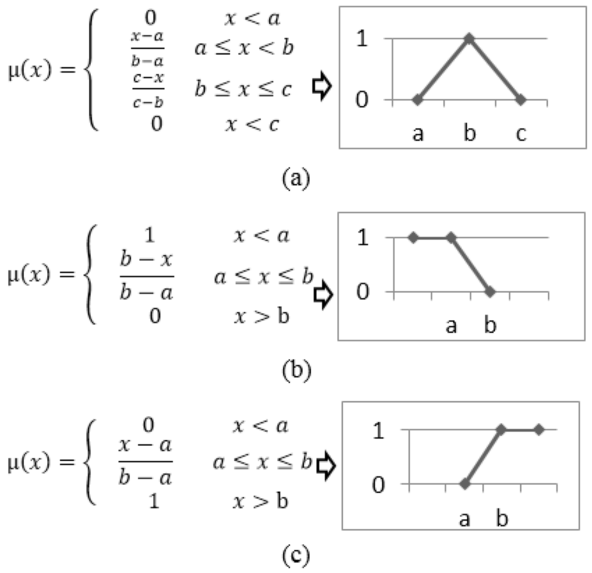
μ(x) = { 0, x < a; (x−a)/(b−a), a ≤ x < b; (c−x)/(c−b), b ≤ x ≤ c; 0, x < c
1
0
a b c
(a)
μ(x) = { 1, x < a; (b − x)/(b − a), a ≤ x ≤ b; 0, x > b
1
0
a b
(b)
μ(x) = { 0, x < a; (x − a)/(b − a), a ≤ x ≤ b; 1, x > b
1
0
a b
(c)

Fig. 2. Three membership functions.

The first condition is the user-defined support approximately 0.8, the word "approximately" (approximately only), it means the function must be a triangulation membership function (For example we used information from the Table I). The values of the variables *a*, *b* and *c* from the triangulation membership function they are meaning, the variable *a* is the minimum value of data, it is 0.1 but to increase the flexibility of the results, in this research will be decreasing the minimum value with minus 0.05 for increase the chances to discover the association rule with the minimum support, therefore the new minimum value is 0.05, the value of the variable *b* is 0.8 because user-defined and the neighboring of this value is the most important than the other values, the variable *c* is the maximum value of data, but to increase the flexibility of the results, in this research will be increasing of the maximum value to add 0.05 to increase the chances to discover the association rule with the maximum support, the new value of variable *c* is 1.05. And then we concluded the value of variables that showed in the Table II and instead support value of each association rule into an equation (Fig. 2 a), after that the results showed in the Table IV (notice: $\mu_A(X)$ From the Table IV if there have variables which has value more than 1 then we will be decreased it to 1, such as there has one variable it has membership value is 1.05 that is more than 1 then we will be decreased it to 1.00).

The second condition is the user-defined confidence approximately more than 0.7, it means the function must be an L membership function because the word "approximately more than", and this function composed two variables *a* and *b* are meaning, the variable *a* is the minimum value of data, the value of the variable *b* is 0.7 because same the reason variable *b* in triangulation membership function, And then We are concluding the value of variables in the Table III and instead confidence of each association rule into an equation (Fig. 2 c), the results showed in the Table IV. In this research does not explain the R membership function because its method is similar to the L membership function.

TABLE II: THE VALUES OF TRIANGULATION MEMBERSHIP FUNCTION

Variable	Support
a	0.05
b	0.8
c	1.05

TABLE III: THE VALUES OF L MEMBERSHIP FUNCTION

Variable	Confidence
a	0.4
b	0.7

TABLE IV: THE RESULTS OF L MEMBERSHIP FUNCTION AND TRIANGULATION MEMBERSHIP FUNCTION

Support	$\mu_A(X)$	Confidence	$\mu_B(X)$
0.1	0.06	0.4	0
0.2	0.2	0.44	0.13
0.3	0.33	0.45	0.16
0.48	0.57	0.56	0.53
0.5	0.6	0.58	0.6
0.6	0.73	0.59	0.63
0.79	~~1.05~~ 1.0	0.62	0.73
0.83	0.88	0.75	1
0.9	0.6	0.76	1
1	0.2	0.8	1

TABLE V: THE SCORES OF EACH ASSOCIATION RULE.

NO.	$\mu_{B(AorB)}(X)$
1	0
2	0.13
3	0.16
4	0.53
5	0.6
6	0.63
7	0.73
8	0.88
9	0.6
10	0.2

D. Computation Score to Rank

This step is the final step to compute the scoring of association rules for select and rank them. Which association rules are selected there score of them must be more than a scoring of user-defined, and in this research we defined the scoring is 0.6. And then we rank the association rules which any association rules have a score more than other rules, they will be ranked in the first order. The computation score with this method, if the user selects the term connected by "and" is used an equation (6), but if the user selects the term connected by "or" is used in equation (7). And then sort the association rules with the scores.

$$\mu_{(AandB)}(X) = \min[\mu_A(X), \mu_B(X)] \quad (6)$$

$$\mu_{(AorB)}(X) = \max[\mu_A(X), \mu_B(X)] \quad (7)$$

From the example 4 with the condition is "and", which can be calculated from the Table IV instead into Equation (6), the results are shown in the Table V and then we selected the association rules with the scoring more than 0.6 the results are No.6, 7, 8 and then we ranked them which any association rules that have a score more than other rules, they will be ranked in the first order, therefore the results are No.8, 7, 6 respectively.

IV. Experiment

This research used data from a random support and confidence (the values are 0.01 to 1.00) to 10k, 50k, 100k and 150k records to test performance of fuzzy searching and to compare non-fuzzy (normal) searching with the various conditions. And the scoring of user-defined in fuzzy searching association rules we defined it to be greater than 0.8. The results are shown in the Table VI.

It can be noticed from the results (Table VI) that the condition can be reduced the number of association rules very much. In particular, the conditions "*s* = 0.52 and *c* =0. 52" because condition "and" to find the minimum value of membership functions, which most association rules do not a predetermined threshold. Therefore, the association rules that have more quality. But the condition "*s* <= 0.52" and condition "*s* >= 0.52" to give similar results because the association rule where the member is 1 will be support begin at the 0.52 (Middle), lead to a similar number of association rules but the most association rules are differently.

TABLE VI: The Results from 5 Conditions of the Fuzzy Searching

Constraint	10k	50k	100k	150k
***s* = 0.52**	2,175	11,042	22,002	32,956
***s* <= 0.52**	6,132	30,613	61,002	91,516
***s* >= 0.52**	6,043	30,431	61,231	91,440
***s* = 0.52 and *c* = 0.52**	468	2,422	4,845	7,225
***s* = 0.52 or *c* = 0.52**	3,876	19,596	39,200	58,743

By the words from Table VI. Represented by these symbols.

"Approximately" represented by "="

"Approximately less than" represented by ">="

"Approximately more than" represented by "<="

"Support" represented by "*s*"

"Confidence" represented by "*c*"

TABLE VII: The Results from 5 Conditions of the Non-Fuzzy Searching

Constraint	10k	50k	100k	150k
***s* = 0.52**	103	490	144	1,485
***s* <= 0.52**	4,906	25,011	75,332	75,016
***s* >= 0.52**	5,197	25,479	76,113	76,469
***s* = 0.52 and *c* = 0.52**	0	0	3	5
***s* = 0.52 or *c* = 0.52**	209	992	285	2,988

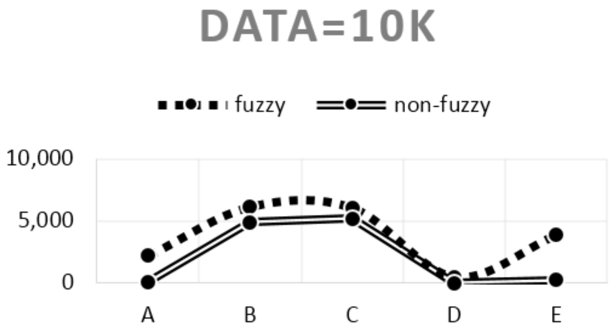

Fig. 3. Fuzzy searching Vs. non-fuzzy (10K).

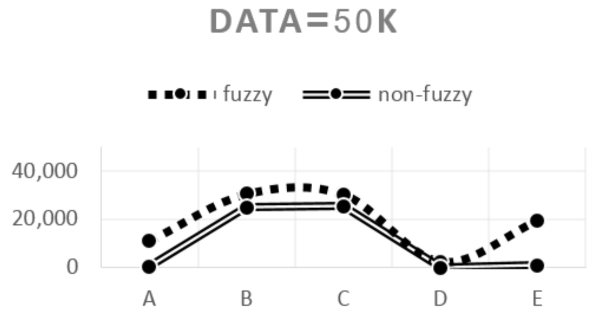

Fig. 4. Fuzzy searching Vs. non-fuzzy (50K).

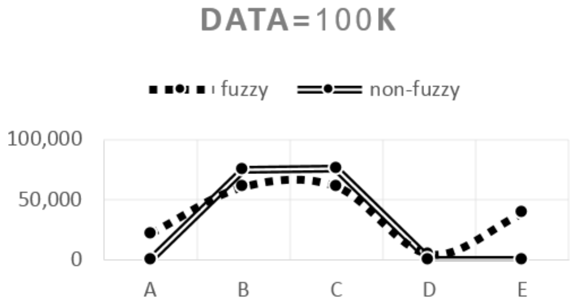

Fig. 5. Fuzzy searching Vs. non-fuzzy (100K).

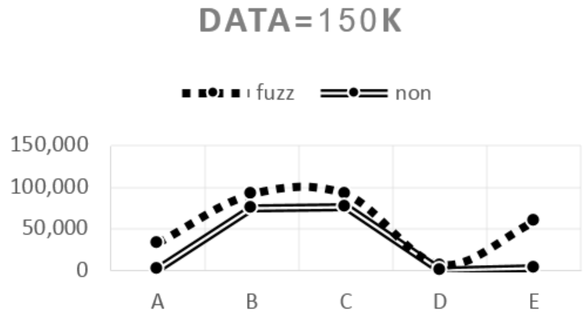

Fig. 6. Fuzzy searching Vs. non-fuzzy (150K).

This results of non-fuzzy (Table VII) is decreased more than fuzzy searching in all cases (Fig. 3 to Fig. 6 Note: the symbols in graphs they mean A: s = 0.52, B: s >= 0.52, C: s = 0.52 and c = 0.52, D: s = 0.52 or c = 0.52), in condition "s <= 0.52" and condition "s >= 0.52" to give a little different form fuzzy searching, and conditions "s = 0.52 and c = 0.52" and conditions "s = 0.52 or c = 0.52" have decreased by almost of 10x. From fuzzy searching, but the condition "s = 0.52 and c = 0.52" do not give the results in data 10k and 50k because don't have association rules are supported = 0.52 and confidence = 0.52, but the fuzzy searching can give the results, such as the association rules have support = 0.51 and confidence = 0.53, support = 0.52 and confidence = 0.53, etc.

Therefore to search the association rule for match the users-requirement should be using non-fuzzy searching, but in some cases this approach cannot provide an answer. However, from that problem should be using the method are flexible and able to give an answer that is close to the most users-requirement this method is fuzzy searching.

V. Conclusion And Future Work

This research proposed methods to search association rules with fuzzy technique from interestingness measures are the support and the confidence. The fuzzy searching are flexible and able to give an answer in some case that is close to the most users-requirement while the non-fuzzy searching do not. In particular, the use of condition "and" can find the rules on stricter conditions than the other conditions then the results are quite similar to the requirements. The conditions "Approximately less than" and "Approximately more than" they give the more results because they are quite flexible conditions. And other conditions the results will be good quality. For future work, we will use this technique to incorporate weight in the association rule discovery with constraint logic to optimize the number of association rules.

References

[1] J. Ha and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2000.

[2] G. I. Webb, "Efficient search for association rules," in *Proc. the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, 2000, pp. 99-107.

[3] B. Jeudy and J. Boulicaut, "Constraint-Based discovery and inductive query: application to association rule mining," *Pattern Detection and Discovery*, pp. 110-124, 2002.

[4] R. T. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang, "Exploratory mining and pruning optimizations of constrained association rules," in *Proc. 1998 ACM SIGMOD Int. Conf. Management of Data*, 1988, pp. 13–24.

[5] R. Srikant, Q. Vu, and R. Agrawal, "Mining association rules with item constraints," in *Proc. the 1997 ACM KDD*, 1977, pp. 67–73.

[6] T. Trifonov and T. Georgieva, "Application for discovering the constraint-based association rules in an archive for unique bulgarian bells," *European Journal of Scientific*, vol. 31, pp. 366-371, 2009.

[7] M. Kawsana and S. Nitsuwat, "Data support system for sales promotion analysis using fuzzy query technique," in *Proc. the 5th National Conference on Computer and Information Technology (NCCIT2009)*, 2009, pp. 684-689.

[8] G. Bordogna and G. Psaila, "Extending SQL with customizable soft selection conditions," in *Proc. the 2005 ACM symposium on Applied computing*, 2005, pp. 1107-1111.

[9] M. Hudec, "Fuzzy improvement of the SQL," in *Proc. the BALCOR 2007 8th Balkan Conference on Operational Research*, 2007, pp. 257-267.

[10] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. the 20th International Conference on Very Large Data Bases*, Santiago, Chile, 1994, pp. 487-499.

[11] W. Siler and J. J. Buckley, *Fuzzy Expert Systems and Fuzzy Reasoning*. Wiley, 2005, ch. 3, pp. 29-54.

Phaichayon Kongchai is currently a doctoral student with the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in computer engineering from Suranaree University of Technology (SUT), Thailand, in 2010, and master degree in computer engineering from SUT in 2012. His current research includes constraint data mining, association mining, functional and logic programming languages, statistical machine learning.

Nittaya Kerdprasop is an associate professor at the School of Computer Engineering, Suranaree University of Technology, Thailand. She received her bachelor degree in radiation techniques from Mahidol University, Thailand, in 1985, master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in computer science from Nova Southeastern University, U.S.A, in 1999. She is a member of ACM and IEEE Computer Society. Her research of interest includes knowledge discovery in databases, artificial intelligence, logic programming, and intelligent databases.

Kittisak Kerdprasop is an associate professor and chair of the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in mathematics from Srinakarinwirot University, Thailand in 1986, master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in computer science from Nova Southeastern University, U.S.A. in 1999. His current research includes data mining, artificial intelligence, functional and logic programming languages, computational statistics.