F.S.A.I., Vol. 8, Nos. 1–3, pp. 25–32, 2002

A Fuzzy Taxonomy-based Method to assess Interestingness of Association Rules

B. Shekar[*] and Rajesh Natarajan

(shek[*], rn)@iimb.ernet.in
Quantitative Methods and Information Systems Area
Indian Institute of Management Bangalore
Bangalore 560076, INDIA

Abstract: We introduce the concept of 'item-relatedness' in the context of association rules (ARs) in Data Mining. ARs containing unrelated or weakly related items are interesting. We elucidate and quantify three different types of item-relatedness. Relationships corresponding to item-relatedness proposed by us are shown to be captured by paths in a 'fuzzy taxonomy.' We then arrive at a total-relatedness measure, demonstrate its efficacy on a sample taxonomy and explain intuitive correspondences between numerical results and reality. Finally, we use this relatedness measure to comment on the interestingness of association rules.
Keywords: Association rules, Fuzzy Taxonomy, Item-relatedness.

1. Introduction

Knowledge discovery in databases (KDD) or Data Mining aims at discovering novel, relevant, significant, valid and useful knowledge from large databases. Association rules [ARs] that bring out co-occurrence of items in a database of transactions are one important class of patterns in KDD. Due to the automated nature of AR discovery, a large number of ARs are extracted the sheer scale of which inhibits comprehension. One approach to tackling this problem involves usage of 'interestingness' measures for ranking ARs. Interestingness measures are classified into objective and subjective categories [1,7]. The approach to evaluating subjective interestingness (of which 'unexpectedness' is a major component) involves assessment of deviations of ARs from user-beliefs [2]. However, this does not take into account, the causes for emergence of unexpectedness.

In our study, we try to capture one aspect of subjective interestingness called 'item-relatedness'. This scheme is based on the idea that ARs that contain unrelated or weakly related items are interesting. Such unrelated item-combinations rarely exhibit high frequencies of occurrence in retail (market basket) transaction databases. Therefore, their presence in ARs is unexpected and hence interesting. In the fuzzy taxonomy (FT) used to represent domain knowledge [3], related items that share common features, occur as leaf-level items in close proximity to each other. This enables the usage of the structural aspects of the FT to derive an intuitive measure of relatedness for item-pairs. The least related item-pair in an AR drives the interestingness of the AR. This pair is identified for each AR and then used to rank the ARs in the increasing order of relatedness.

Here, we consider the retail market-basket context.

A retail database consists of items purchased by a customer on a per transaction basis. We assume that ARs have been discovered and a FT, that incorporates domain knowledge, is available. Figure 1 gives an overview of the approach adopted here. This study specifically focuses on the development of a relatedness measure that can be used for ranking ARs that are inherently non-fuzzy [5]. The feedback in the Figure 1 indicates that newly discovered knowledge might influence the beliefs of the user. This, in turn, can lead to modifications in the structure of the FT. Here, we use standard fuzzy set related [4] and graph-theoretic [6] terminologies.

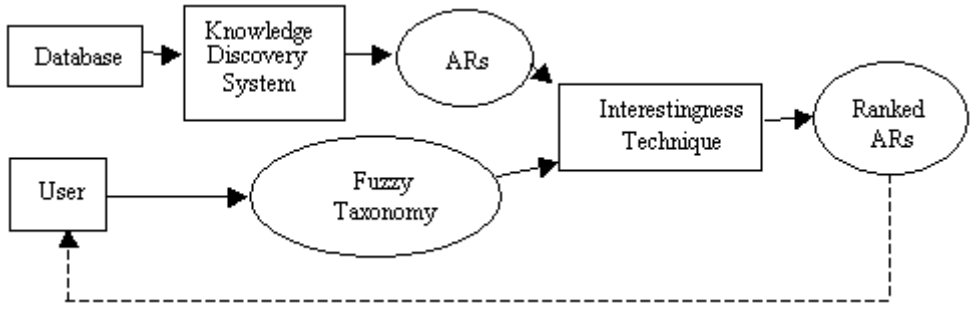

Fig. 1 – An approach to identification of interesting patterns

2. Fuzzy Taxonomy (FT)

A Fuzzy Taxonomy (FT), an extension of the traditional concept hierarchy tree (is-a relationship), enables us to express fuzzy (incomplete) relationships existing between items and their higher-level concepts [3]. This enables categorization of items that might belong to more than one higher-level concepts. A FT is a tree with items represented as leaf nodes and concepts represented as non-leaf nodes. A membership function (μ) gives the extent to which a 'child' node belongs to its 'parent' category. This function takes a value between 0 and 1. The membership value is then carried over to other ancestors (emanating from its parent) of the 'child' node. When an 'ancestor' node has two or more membership grades from the various parents of a leaf-level item, we choose the highest membership value. This is to preserve the membership value of the leaf-level item in its closest ancestor. Formally, the membership function of child node '*c*' in its parent node '*p*' is given by: $\mu_{(c,p)} = x : 0 \leq x \leq 1$. For ancestor '*a*', the membership grade is given by $\mu_{(c,a)} = \max\{ \mu_{(c,a)} (p)\}$ where $p = 1, .., N$; and $\mu_{(c,a)} (p)$ is the membership grade of child node '*c*' in its ancestor '*a*' by virtue of path '*p*'. In a very broad sense, transfer of memberships from the child to parent nodes, gives an indication of the extent to which properties of the child overlap with those of the parent node.

Highest-level node of path [$H_{A, B}$ (p)]: The highest-level node of path(*A*, *B*) is defined as the node that occurs at the highest level (i.e. nearest to the root node) in

the path *p* connecting items *A* and *B*. The highest-level node of a path is the closest common context that relates two items. Therefore, the distance of the highest-level node from the root node gives an indication of the relatedness between the two items. A larger distance implies greater relatedness. As we move down the taxonomy from the highest-level node to the two items, the differences between the two items gain prominence.

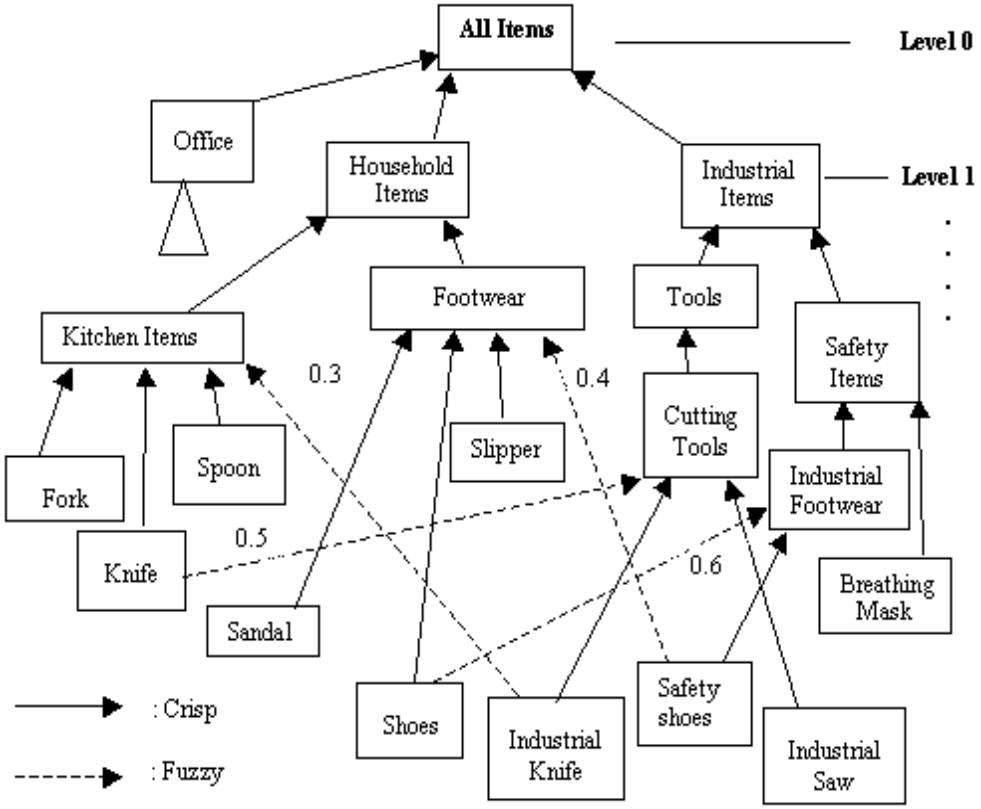

Fig. 2 – A Fuzzy Taxonomy

3. Item Relatedness

Relationships between items are a consequence of the generic (primary functional purpose) category of items, domains of application and secondary functionalities. Any two items, say *A* and *B*, can be connected by many such relationships of varying strengths. All ARs have support (statistical significance) and confidence (predictive ability) greater than user specified minimum values. An item-pair in an association rule is 'interesting' to a user if the items in it are weakly related to each other. High frequency of occurrence of these weakly related pairs

makes them interesting due to their unexpectedness. Thus, 'relatedness' and 'interestingness' are opposing notions. For example, an item pair {beer, diaper} will evoke interest because they are weakly related. These items have generically different functions. They are used in different domains of application. On the other hand, item pair {bread, butter} will evoke little interest, as these are highly related items. We now discuss different types of relatedness.

Highest-level Node Membership [$HM_{A,B}(p)$]: The fuzzy path that connects any two items in a FT indicates partial membership of the items in each one of the nodes lying in the path. Highest-level node membership of items *A* and *B* in path '*p*' is given by the minimum of the membership values of the two items in the highest-level node of path '*p*':

$$HM_{A,B}(p) = \min\left[\mu_{A,H(A,B)}(p), \mu_{B,H(A,B)}(p)\right]$$

For example, the node **Industrial items** is the highest-level node of the fuzzy path connecting items **knife** and **shoes** [Figure 2]. Thus,

$$HM_{(knife,\ shoes)}(p) = \min[0.5, 0.6] = 0.5.$$

'Minimum' is used as the operator here because it effectively conveys the maximum extent to which one item can substitute another in the same usage context. Thus, **knife** and **shoes** cannot be considered to be proper **Industrial items**, but they can be used together as 'industrial items' to a maximum extent of 0.5. On the other hand, the use of some other operator to link the two items, say average (whose value is 0.55), overstates the membership of **Knife** in **Industrial Items.**

Highest-level Relatedness [$HR_{A,B}(p)$]: The distance of the "highest-level node" [$H_{A,B}(p)$] from the root node gives another component of relatedness. Nodes close to the root node deal with attributes that reveal the separation of the entire domain of items into various sub-domains. These attributes tend to be fundamental (or basic) as far as characterization of the domain is concerned. Their values segregate and categorize items into various categories/sub-categories. Closer the highest-level node (of path '*p*') to the root node, greater is the degree of separation between the items and consequently lower is the relatedness. We capture this notion by a measure called 'Highest-level relatedness' [$HR_{A,B}(p)$] which is the level of the highest-level node [$H_{A,B}(p)$] in path '*p*'.

$$HR_{A,B}(p) = level\left[H_{A,B}(p)\right]$$

Node Separation Relatedness [$NSR_{A,B}(p)$]: The length of the path (in terms of nodes) connecting items *A* and *B* in a fuzzy taxonomy gives an indication of the conceptual distance between them. Category nodes other than the highest-level node might consider attributes that are instantiated and specific to one of the two items under consideration. A large number of such distinct attributes specific to each item (resulting in a longer path) reduce relatedness between items. We define a measure of item-relatedness called node-separation relatedness, $NSR_{A,B}(p)$, which is the length of the simple path [6] '*p*' connecting *A* and *B*.

$NSR_{A,B}$ (p) = Length of the simple path 'p' (in terms of number of nodes) connecting nodes A and B.

4. A Fuzzy Item-Relatedness Measure and an Illustration:

Consider two items, say **knife** and **shoes** from Figure 2. We find that four components of relatedness (represented by the four paths) exist between them. Relatedness between items *A* and *B* increases if there is an increase in either highest-level node membership (HM) or highest-level relatedness (HR) or both. On the contrary a decrease in node-separation relatedness measure (NSR) increases relatedness. Overall relatedness contributed by a single path 'p' can be given by:

$$\mathbf{OR_{A,B}(p)} = \frac{\mathbf{(1 + HR_{A,B}(p))(HM_{A,B}(p))}}{\mathbf{NSR_{A,B}(p)}} \quad (1)$$

The highest-level relatedness is incremented by 1 to account for the case where the root node appears as highest-level node. It can be easily shown that for a FT of depth '*k*', the maximum relatedness value contributed by a single path between any two leaf-level items is '*k*'. This value can be used to normalize the overall relatedness value. Consequently, total relatedness (a summation over all paths) can be given by:

$$\mathbf{TR(A,B)} = \sum_{\mathbf{p}} \frac{\mathbf{OR_{A,B}(p)}}{k} = \sum_{\mathbf{p}} \frac{\mathbf{(1 + HR_{A,B}(p))(HM_{A,B}(p))}}{\mathbf{k \times NSR_{A,B}(p)}} \quad (2)$$

Table 1 shows the 'relatedness' components ($OR_{A,B}(p)$ and $TR(A,B)$) for five pairs of items from Figure 2. Let us consider item pair {**shoes**, **safety shoes**}. The Crisp path I consider **shoes** as household item and **safety shoes** as an industrial item while path II (fuzzy path) considers the reverse. The relatedness components contributed by these two paths are quite low. This is expected as these paths consider relationships when the two items are in different domains namely household and industrial domains.

The fuzzy path contribution is lower than the crisp path contribution as the items are not 'complete' members in their respective domains. Path IV (crisp-fuzzy path) considers both **shoes** and **safety shoes** as **footwear** items under **household items**. Although the membership of **safety shoes** in **footwear** is fuzzy (0.4), the fact that the two items are siblings and thus can substitute one another in the same parent domain (**household**) to a maximum extent of 0.4 strengthens the relatedness between **shoes** and **safety shoes**.

This feature is reflected by a high contribution (0.30) to the total relatedness measure. Similarly, **safety shoes** and **shoes** when used in industrial domain have a relatedness of 0.60. Intuitively, we know that **Shoes** and **Safety Shoes** is footwear whose primary purpose is to protect the feet of the wearer. Naturally, we would expect a high relatedness between them, except for the fact that they are normally

used in different scenarios. Paths I and II emphasize the fact that they belong to different domains. Failure to consider every path will result in an understatement of the total relatedness.

The relatedness of item pairs 3 and 4 show that items belonging to the same sibling domain are related to a greater extent than items belonging to differing sibling domains. We can see from item pairs 2 and 4 that relatedness between two items is strengthened if greater numbers of relationships exist between them.

Table 1

A comparison of Item Relatedness for sample item-pairs

Sr. No.	Item A	Item B	Path I (Crisp-Crisp)	Path II (Fuzzy-Fuzzy)	Path III (Fuzzy-Crisp)	Path IV (Crisp-Fuzzy)	TR(A, B)
1.	Knife	Shoes	0.1667	0.0500	0.0208	0.0250	0.2625
2.	Shoes	Safety Shoes	0.0417	0.01667	0.6000	0.3000	0.9584
3.	Spoon	Shoes	0.1667	----	----	0.0250	0.19175
4.	Spoon	Safety Shoes	0.04167	----	----	0.0667	0.10835
5.	Knife	Industrial Knife	0.04168	0.0125	0.5000	0.225	0.7792

5. An Application in a Retail Market-Basket Scenario

ARs indicate high frequencies of repeated purchases by customers. ARs composed of items that are unrelated in normal circumstances are interesting. This is because such rules are a consequence of some interesting phenomenon taking place in the background. A user may be interested in such rules because he/she may be unaware of them. Relatedness values derived from the structure of the FT using Equation 2 help in the identification of such interesting ARs. Leaf nodes in close proximity represent related items. Presence of a large number of strong relationships between items (represented by normal and fuzzy paths) increases relatedness.

When a user examines an AR, the least related item-pair in the AR, is the item-pair that stands out. This is because such item-pairs are not expected to be frequent in the retail transaction database. Thus, the least related item-pair of an AR drives the interestingness of the AR. Hence, we identify the least related item-pair in an AR and assign its relatedness value to the rule. This gives a rough estimate of the rule's interestingness. Rules are ranked according to an ascending order of their relatedness. The rule having the smallest relatedness value is then the most interesting one.

Let us assume that we have a knowledge base given by the FT of Figure 2. Let us also assume that the sample ARs (shown in Table 2) have been generated from a retail database of purchase transactions. These ARs have the required 'support' and 'confidence'[7]. We have ranked these ARs in the ascending order of their TR values. We notice that Rule 2 has a larger relatedness value than Rule 1.

Table 2

Ranking of Sample Association Rules

Sr. No.	Rank	Association Rule (AR)	Relatedness (TR)
1.	1	{Breathing Mask, Industrial Saw}⇒{Spoon}	0.04167
2.	2	{Breathing Mask, Industrial Saw}⇒{Knife}	0.11250
3.	3	{Sandals, Slippers}⇒{Fork}	0.16667
4.	4	{Sandals, Slippers}⇒{Safety Shoes}	0.34167
5.	5	{Knife, Fork}⇒{Spoon}	0.75000

This is because partial utility of household item **knife** in the industrial domain increases its relatedness to other items of the rule. Note that Rules 1 and 2 have the same industrial items in their antecedent. It is interesting to note that even though all items of Rule 3 are from the 'household' domain (and therefore its items are expected to be more related), this rule is more interesting than Rule 4. Rule 4 has two items from the 'household' domain and one from the 'industrial' domain. This deviation from expectation occurs because item **safety shoes** (an item from the industrial domain) of Rule 4 can also be used in the household domain to a limited extent. In addition, note that all three items of Rule 4 belong to the category **footwear** (and thus have the same primary purpose) though in different domains. Finally, we note that Rule 5 is ranked the lowest on interestingness as it consists of three **kitchen items**. One can expect them to be purchased together. Thus, we see that TR-measure-based ranking of ARs is intuitively appealing.

6. Summary and Conclusions

AR studies [1,2,7] have frequently pointed out problems associated with generation of numerous irrelevant and obvious rules. Here, we have used the notion of 'item relatedness' to capture one aspect of subjective interestingness. Association rules that contain unrelated or weakly related combinations are the ones that are interesting, because frequent occurrences of such item combinations are rare. We have derived an intuitive measure of relatedness using the structure of a FT. An item pair has a large relatedness value if its constituent items occur together in close proximity in the FT and if there are a large number of strong relationships between them. In addition to bringing out the appropriateness and intuitiveness of the measure, we have also demonstrated the efficacy of the relatedness measure in ranking rules based on their interestingness. FT forms the user knowledge base. It can be iteratively evolved over a period of time. Use of FT

helps in the knowledge elicitation process as users can now be guided by its structure. Further, the universality of the categorization process ensures that some objectivity is ensured in the elicitation process. On the other hand, unstructured knowledge elicitation is incomplete and difficult. 'Item-Relatedness' as a concept is useful for identifying interesting rules. It uses domain knowledge of users in the form of a FT during the identification process. The FT represents the 'is-a' relationships between items and their categories. It can take into account the functional separation between items. Therefore, AR rankings obtained with the help of FT are intuitive. Our work is a small step aimed at furthering the understanding of 'Interestingness' in the context of Association Rule Mining.

References

[1] A. SILBERSCHATZ, A. TUZHILIN: *What makes Patterns Interesting in Knowledge Discovery Systems*, IEEE Transactions on Knowledge and Data Engineering, Vol. **8** (6), 970-974, 1996.

[2] B. LIU, W. HSU, S. CHEN, Y. MA: *Analyzing the Subjective Interestingness of Association Rules*, IEEE Intelligent Systems, vol. **15** (5), 47-55, 2000.

[3] G. CHEN, G. WETS, K. VANHOOF: *Representation and Discovery of Fuzzy Association Rules (FARs)*, Institute of applied Economic Research (ITEO), Limburg University Center (LUC), Belgium, Research Paper Series, ITEO No: 00/01, March 2000.

[4] G. J. KLIR, B.YUAN: *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, Prentice Hall of India Private Limited, New Delhi, 1997.

[5] J.M. DE GRAAF, W.A. KOSTERS, J.J.W. WITTEMAN: *Interesting Fuzzy Association Rules in Quantitative Databases*, in Proc. of PKDD 2001 (The 5th European Conference on Principles of Data Mining and Knowledge Discovery), Springer Lecture Notes in Artificial Intelligence 2168, editors L. De Raedt and A. Siebes, pp. 140-151, Freiburg, Germany, September 3/5, 2001.

[6] N. DEO: *Graph Theory with Applications to Engineering and Computer Science*, Prentice Hall of India Private Limited, 1989.

[7] P. TAN, V. KUMAR, J. SRIVASTAVA: *Selecting the Right Interestingness Measure for Association Patterns*, accepted for the Eighth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD-2002), July 23-26, 2002.