A relatedness-based data-driven approach to determination of interestingness of association rules

Conference Paper · January 2005

DOI: 10.1145/1066677.1066803 · Source: DBLP

2 authors:

Rajesh Natarajan
University Carlos III de Madrid
18 PUBLICATIONS **92** CITATIONS
SEE PROFILE

B. Shekar
Indian Institute of Management Bangalore
31 PUBLICATIONS **168** CITATIONS
SEE PROFILE

A Relatedness-based Data-driven Approach to Determination of Interestingness of Association Rules

Rajesh Natarajan
Information Technology and Systems Group
Indian Institute of Management Lucknow
Lucknow – 226 013, INDIA.
91-522-273 4101
rajeshn@iiml.ac.in

B. Shekar
Quantitative Methods and Information Systems Area
Indian Institute of Management Bangalore
Bangalore – 560 076, INDIA.
91-80-2699 3093
shek@iimb.ernet.in

ABSTRACT
The presence of unrelated or weakly related item-pairs can help in identifying Interesting Association Rules (ARs) in a market basket. We introduce three measures for capturing the extent of mutual interaction, substitutive and complementary relationships between two items. Item-relatedness, a composite of these relationships, can help to rank interestingness of an AR. The approach presented, is intuitive and can complement and enhance classical objective measures of interestingness.

Categories and Subject Descriptors
H.2.8 [**Information Systems**]: Database management– *database applications- data mining.* H.4.2 [**Information Systems**]: Information Systems Applications– *types of systems – decision support.*

General Terms
Management, Measurement, Human Factors

Keywords
Data mining, Association rules, Interestingness, Relatedness.

1. INTRODUCTION
Ranking of ARs based on interestingness is an important approach to addressing the rule immensity problem in AR mining. Interestingness measures quantify the amount of interest that a rule is expected to evoke on inspection. Interestingness, an elusive concept, may be objective [1, 4] or subjective [2, 3]. The presence of unrelated or weakly related items in an AR can make it interesting. Relatedness is a consequence of relationships that exist between items in a domain. In addition to co-occurrence, an examination of other relationships such as complementarity and substitutability will reveal relatedness between items in a more intuitive fashion. If we consider complementary relationships that bind items, then the relatedness of {bread, butter} becomes higher than {beer, diaper}. Measures based on co-occurrence alone may not reveal this.

Our approach to interestingness is objective and data-driven because relatedness is determined solely on the examination of purchase transactions without taking recourse to domain knowledge. Relatedness between two items is deduced from the co-occurrence frequency and associated purchases. The opposing nature between relatedness and interestingness can be used to quantify interestingness of ARs.

SAC'05, March 13-17, 2005, Santa Fe, New Mexico, USA.

2. ITEM RELATEDNESS
Relationships originate from interaction between functions of items. Item-relatedness is a composite of these relationships. Usage of items in a domain dictates their purchase. Hence purchase transactions can indicate relatedness.

Consider an item-pair {x, y}. Let t_x denote the set of transactions that contains item x but not item y. 't_y' is similarly defined. t_{xy} denotes the set that contains both x and y. The transactions in these sets may contain items other than x and y accordingly. Items purchased with x and y either together or individually can contribute substantially to the relatedness between them. This is because usefulness of {x, y} could increase when used with them. The co-occurring neighbourhood Z of {x, y} is the set of items other than x and y, which occurs in t_{xy} in significant numbers. The non co-occurring neighbourhood $M \cap N$ is the overlap of item-sets M (= t_x – x) and N (= t_y – y**).** Note that items should have a significant presence in t_x and t_y to qualify as members of M and N respectively**.** An examination of t_{xy}, Z and ($M \cap N$) can reveal relationships that bind x and y.

If x and y together imply a useful function then they complement each other. Generally, complementary items occur in the same transaction. Hence, a high (significant) cardinality t_{xy} ($|t_{xy}|$) reveals complementarity. Items co-purchased with {x, y} reveal various shades of complementarity. If $Z=\phi$, then complementarity is 'intrinsic'. If $Z\neq\phi$, then complementarity is 'dependent'. This is because each $z \in Z$ along with {x, y} serves a useful function.

If x serves y's function to a significant extent then x substitutes y. Substitutes have similar/closely related properties. Substitutes are

not likely to occur in a single transaction. Hence $|t_{xy}|$ will not be significant. However, they may be purchased in t_x and t_y with similar items, $M \cap N \neq \phi$. If $|t_{xw}|$ ($t_{xw} \subset t_x$) and $|t_{yw}|$ ($t_{yw} \subset t_y$) are significant, then x and y may be deemed substitutes in w's presence.

Two items are non-dependent if they unrelated through either of mutual interaction, substitutability and complementarity. For two non-dependent items, $(M \cap N) \cup Z = \phi$ and $|t_{xy}|$ is insignificant. Note that complementarity and substitutability are not mutually exclusive notions.

3. MEASURES FOR RELATEDNESS

Mutual interaction between two items {x, y}, is captured by R_1:

$$R_1 = 0.5 \times \left[\frac{f(xy)}{f(x)} + \frac{f(xy)}{f(y)} \right] \quad (1)$$

where, f(x) is the frequency of item x (i.e. $|t_x| + |t_{xy}|$) while f(xy) is the frequency of transactions that contain both x and y (i.e. $|t_{xy}|$). R_1, a modification of confidence, is the average predictive ability of the presence of one item given the presence of the other. R_1 ignores the presence of other items in t_x, t_y, and t_{xy}.

R_2 reveals the intensity of dependent complementarity, based on the co-occurring neighbourhood of {x, y}. Items that co-occur with x and y can point to the varied usage of item-pair {x, y}.

$R_2 = 0$ if $|Z|=0$ i.e. $Z=\phi$

$$= \frac{1}{|Z|} \times \sum_{z \in Z} \left[\frac{f(xyz)}{f(xy)} \right] \quad \text{(otherwise)} \quad (2)$$

$z \in Z$ only if $|t_{xyz}|$ is significant, i.e. $\left[\frac{f(xyz)}{f(xy)} \right] \geq \min\ sup$. Every term in the summation has a value in the range [0, 1] indicative of the strength of dependent complementarity related to each z.

R_3 uses the non co-occurring neighbourhood of {x, y} to compute the degree of substitutability.

$R_3 = 0$ if $(M \cap N) = \phi$

$$= \frac{1}{|(M \cap N)|} \times \sum_{w \in (M \cap N)} \left[1 - |a_w - b_w| \right] \quad \text{(otherwise)} \quad (3)$$

where $a_w = \left(\frac{f(xw) - f(xyw)}{f(x) - f(xy)} \right)$ and $ß_w = \left(\frac{f(yw) - f(xyw)}{f(y) - f(xy)} \right)$ and $\alpha_w, \beta_w \geq Sig$, a significance threshold.

$|\alpha_w - \beta_w|$ gives the deviation of the proportionate occurrence of w with x and y. Strong substitutes will be similar to each other with respect to their non co-occurring neighbours i.e. $\alpha_w \approx \beta_w$.

R1, R2 and R3 vary in the range [0, 1]. The value of each measure gives the strength of one relationship. Hence, relatedness of two items can be quantified by Total Relatedness: TR=R1+R2+R3

4. DISCUSSIONS AND CONCLUSIONS

It can be intuitively argued that the least related item pair of an AR drives its interestingness [2]. In addition, an AR that contains a larger number of items will have greater interestingness. These notions, along with the inverse nature of relationship between interestingness and relatedness, can be used to develop interestingness coefficients for an AR as in [2].

An observation on TR is as follows. Although any one relationship can dominate the interaction, the total relatedness is correct only if we consider all relationships. This is because TR depends on both, cardinality of relationships and strength of each relationship. Therefore, two items {x, y} strongly related only through co-occurrence may be less related than another pair {a, b} exhibiting relatively weaker presence in all three relationships. Note that a single relationship can contribute a maximum value of 1. A stronger relationship (TR) should have more components. This essentially means a pair if used with a larger number of items in more situations, is more related and hence less interesting from a usage point of view.

Since TR does not consider the implication sign, ARs having identical sets of items will have the same interestingness values. This situation can be handled by a two-stage ranking process. Interestingness coefficients based on TR can be used to prune out definitely uninteresting rules. Classical objective measures that consider implication, like conviction and confidence can then be used in the second stage for further discernment. Alternatively, the sets of item-pairs in the antecedent and consequent can be considered separately. Item-pairs in the antecedent, consequent, and antecedent-consequent (pairs formed with one item from each), may be weighed differently in accordance with the relative importance ascribed to them and then combined accordingly for the whole AR. Note that rules selected by redundancy reduction methods can be the input set for the proposed scheme. Hence, redundancy reduction methods complement interestingness-based ranking schemes.

We compared our AR ranking scheme with 'rule interest' and 'conviction' [4] using a sample artificial dataset. Interestingness rankings based on TR were more intuitive as TR considers other aspects of an item-pair's relatedness in addition to co-occurrence. Future work is in the direction of extending TR beyond item-pairs, incorporating directional aspects of implications into interestingness evaluation, and testing on real-life datasets. The work reported here is one approach to interestingness, based on data-driven relatedness.

5. REFERENCES

[1] Omiecinski, E.R. Alternative Interest Measures for mining Associations in Databases. *IEEE Transactions on Knowledge and Data Engineering*, 15, 1(Jan/Feb 2003), 57-69.

[2] Shekar, B. and Natarajan, R. A Framework for Evaluating Knowledge-based Interestingness of Association Rules. *Fuzzy Optimization and Decision Making*, 3, 2 (June 2004), 157-185.

[3] Silberschatz, A. and Tuzhilin, A. What makes Patterns Interesting in Knowledge Discovery Systems. *IEEE Transactions on Knowledge and Data Engineering*, 8, 6(1996), 970-974.

[4] Tan, P., Kumar, V. and Srivastava. J. Selecting the Right Interestingness Measure for Association Patterns. Information Systems, 29, 4 (June 2004), 293-331.