IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITB12610

This paper appears in the book, *Emerging Trends and Challenges in Information Technology Management, Volume 1 and Volume 2* edited by Mehdi Khosrow-Pour © 2006, Idea Group Inc.

Weakness of Association Rules: A Mechanism for Clustering

Rajesh Natarajan, IT & Systems Group, Indian Institute of Management Lucknow, Lucknow - 226 013, Uttar Pradesh, India, T: +91-522-2736659, F: +91-522-2734025, rajeshn@iiml.ac.in

B. Shekar, Quantitative Methods & Information Systems Area, Indian Institute of Management Bangalore, Bangalore - 560 076, Karnataka, India, T: +91-80-26993093, F: +91-80-26584050, shek@iimb.ernet.in

ABSTRACT

We introduce the notion of *weakness* of an AR. After providing the intuition, we develop a *weakness-based* distance-function for clustering ARs. We cluster ARs obtained from a small artificial data set through the average-linkage method. The clusters are compared with those obtained by applying a commonly used method to the same data-set.

1. INTRODUCTION

Rule immensity is an important issue in Association Rule (AR) mining. This problem concerns the multitude of discovered rules that hinder easy comprehension. We define *Weakness* as the extent to which an AR is unable to explain the presence of its constituent items. Weakness is then used as a heuristic to group ARs. Rules with similar *weakness* are placed in the same cluster, thus facilitating easy exploration of connections among them. A user needs to examine only those rules in 'relevant' clusters.

Lent, Swami and Widom [6] introduced the notion of 'clustered' ARs. Adomavicius and Tuzhilin [1] adopted an expert-driven, attribute hierarchy-based similar rule-grouping approach. The distance measure proposed by Toivonen, et al. [8] and Gupta and others [3] clustered rules that 'cover' the same set of transactions. One limitation of [8,3] is the arbitrariness of distance measures [1].

Dong and Li [2] introduced a distance metric for detecting unexpected rules. Sahar's d_{SC} [7] utilized both syntactic matching of item-sets and rule coverage of data. Jorge [5] studied clustering in the context of thematic browsing and summarization of large sets of ARs. Current research has concentrated either on syntactic (item–matching based) comparison [1,2,5] or on transaction-set coverage [3,7,8]. These approaches do not utilize certain intrinsic properties of ARs. We propose *weakness* (an intrinsic property)–based identification of specificity/generality of the AR in describing the presence of its constituents in the database.

2. WEAKNESS OF AN ASSOCIATION RULE

Consider an AR, $R: a_1a_2..a_m \rightarrow a_{m+1}a_{m+2}\ldots a_n$, having support S_R and confidence C_R. If all items of R are present in that transaction (t), then R covers t. Let the support of an individual item $a_i \in R$ with respect to database D be Sa_i. R accounts for only S_R% of transactions in the database and does not explain the portion (of D) containing $1 - \frac{S_R}{S_{a_i}}$ % of transactions containing a_i. This fraction may be viewed as *weakness* of R with respect to its constituent a_i: $w_{a_i} = 1 - \frac{S_R}{S_{a_i}}$ (1)

Weakness of an AR with respect to all its constituents is given by:

$$w_R = \frac{1}{n}\sum_{a_i}\left(1 - \frac{S_R}{S_{a_i}}\right);\ a_i \in \{a_1, a_2, ..., a_n\} \quad (2)$$

'w-value' brings out the strength of relationship between an AR and its constituents. A low w-value indicates strong characterization of its constituent items, since most of the transactions containing R's constituent items exhibit the behaviour captured by R. In addition, a low w-value signifies generality (wider coverage in D) of the relationship described by R. In contrast, a high w-value indicates specificity of the relationships revealed by the rule.

3. A WEAKNESS-BASED DISTANCE MEASURE (d_w)

Low generality of a high w-value rule suggests that relationships between the rule's items and items present in other rules may be worth exploring. Actions taken only on the basis of a high w-value (high-specificity) rule could be skewed as the rule brings out only one aspect of the items' behaviour in the database. Since *weakness* reflects the presence of relationships among constituents, action based on rules with equal or near-equal values could yield similar results.

We define *weakness*-based distance as:

$$d_w(R_1, R_2) = \frac{|w_1 - w_2|}{w_1 + w_2},\ 0 \le w_1, w_2 \le 1. \quad (3)$$

Any difference Δw results in a larger distance for low w-values and smaller distance for high w-values. *If* $(|w_1\text{-}w_2|=|w_3\text{-}w_4|)$ *and* $(w_1+w_2 \le w_3+w_4)$*, then* $d_w(R_1,R_2) > d_w(R_3,R_4)$. Let w_1=0.4, w_2=0.2, w_3=0.8 and w_4=0.6. Then, $d_w(R_1,R_2)=0.3333$ while $d_w(R_3,R_4)=0.14285$. This may seem counter intuitive. However it has a rationale. R_1 and R_2 are unable to explain 40% and 20% respectively of their constituent items' presence. Thus, they are more *general* than R_3 and R_4 whose w-values are 0.8 and 0.6 respectively. R_3 and R_4 have poorer explanatory power than R_1 and R_2, with respect to their constituent items.

This rationale has an analogical intuitive support. Consider four individuals $A(R_1)$, $B(R_2)$, $C(R_3)$ and $D(R_4)$. Assume A and B possess deeper knowledge (of a topic) than C and D. Let the absolute difference in the knowledge-levels between the individuals in each of $\{A,B\}$ and $\{C,D\}$ be the same. Since A and B are quite knowledgeable, the difference would seem to be larger because it would require more effort to move from A's knowledge-level to B's knowledge-level. This greater effort may be due to the subtle and conceptually deeper knowledge required. However, it may be relatively easier to bridge the gap between C and D. Fewer facts and straightforward knowledge acquisition may suffice. Similarly, R_1 and R_2 may have good explanatory power and hence they may be separated by a larger distance than the more specific pair $\{R_3,R_4\}$.

Table 1. An artificial transaction dataset

Transaction	Nos.	Transaction	Nos.
{Bread,Butter}	6	{Bread,Jam}	5
{Bread,Milk}	4	{Bread,Butter,Milk}	10
{Milk,Chocolate}	6	{Chocolate,Biscuit}	8
{Milk,Chocolate,Biscuit}	11	{Butter,Milk}	3
{Pen,Pencil,Eraser}	13	{Pencil,Eraser}	7
{Chocolate,Pencil,Eraser}	3	{Pen,Eraser}	3
{Chocolate,Biscuit,Pencil}	5	{Bread,Butter,Milk,Jam}	4
{Bread,Jam,Milk}	12	--	--

Table 2. Association Rules extracted from transaction set of Table 1

No	Rule	Support	Confidence	Weakness
R_1	Butter→Bread	0.20	0.86957	0.321315
R_2	Jam→Bread	0.21	1.00	0.243902
R_3	Bread→Milk	0.30	0.7317	0.334146
R_4	Butter→Milk	0.17	0.73913	0.460435
R_5	Butter,Milk→Bread	0.14	0.82353	0.589947
R_6	Chocolate→ Biscuit	0.24	0.72727	0.136364
R_7	Milk,Biscuit→Chocolate	0.11	1.00	0.662778
R_8	Pen→Pencil,Eraser	0.13	0.8125	0.407738
R_9	Pen→Pencil	0.13	0.8125	0.361607
R_{10}	Pencil→Eraser	0.23	0.82143	0.146978
R_{11}	Pen→Eraser	0.16	1.00	0.192308
R_{12}	Jam,Milk→ Bread	0.16	1.00	0.509284
R_{13}	Jam→Milk	0.16	0.76190	0.459048
R_{14}	Chocolate→ Milk	0.17	0.51515	0.572424

Table 3. d_w –based clustering

Step_No	Clusters
1	$\{\mathbf{R_{13},R_4}\}$ [0.002]
2	$\{\mathbf{R_{14},R_5}\};\{R_{13},R_4\}$ [0.015]
3	$\{\mathbf{R_3,R_1}\};\{R_{14},R_5\};\{R_{13},R_4\}$ [0.020]
4	$\{\mathbf{R_{10},R_6}\};\{R_3,R_1\};\{R_{14},R_5\};\{R_{13},R_4\}$ [0.037]
5	$\{\mathbf{R_3,R_1,R_9}\};\{R_{10},R_6\};\{R_{14},R_5\};\{R_{13},R_4\}$ [0.049]
6	$\{\mathbf{R_{13},R_4,R_{12}}\};\{R_3,R_1,R_9\};\{R_{10},R_6\};\{R_{14},R_5\}$ [0.051]
7	$\{\mathbf{R_{14},R_5,R_7}\};\{R_{13},R_4,R_{12}\};\{R_3,R_1,R_9\};\{R_{10},R_6\}$ [0.066]
8	$\{\mathbf{R_8,R_{13},R_4,R_{12}}\};\{R_{14},R_5,R_7\};\{R_3,R_1,R_9\};\{R_{10},R_6\}$ [0.077]
9	$\{\mathbf{R_{11},R_2}\};\{R_8,R_{13},R_4,R_{12}\};\{R_{14},R_5,R_7\};\{R_3,R_1,R_9\};\{R_{10},R_6\}$ [0.118]
10	$\{\mathbf{R_8,R_{13},R_4,R_{12},R_{14},R_5,R_7}\};\{R_{11},R_2\};\{R_3,R_1,R_9\};\{R_{10},R_6\}$ [0.140]
11	$\{\mathbf{R_8,R_{13},R_4,R_{12},R_{14},R_5,R_7,R_3,R_1,R_9}\};\{R_{11},R_2\};\{R_{10},R_6\}$ [0.207]
12	$\{\mathbf{R_{11},R_2,R_{10},R_6}\};\{R_8,R_{13},R_4,R_{12},R_{14},R_5,R_7,R_3,R_1,R_9\}$ [0.209]
13	$\{\mathbf{R_{11},R_2,R_{10},R_6,R_8,R_{13},R_4,R_{12},R_{14},R_5,R_7,R_3,R_1,R_9}\}$ [0.435]

Note: Values in the brackets represent merging distance

It is easy to establish the metric properties of $d_w(R_i,R_j)$. The intuitive justification of $d_w(R_i,R_j)$ and its being a metric enable d_w-based clustering of ARs.

4. d_w-BASED CLUSTERING OF ARs

Table 1 represents an artificial transaction database consisting of 100 transactions; the complete item-set being {*Bread*,*Butter*,*Jam*,*Milk*, *Chocolate*,*Biscuit*,*Pen*,*Pencil*, *Eraser*}. It contains fifteen unique market-baskets. Support and confidence having respective thresholds of 0.1 and 0.5 yielded fourteen ARs listed in Table 2.

R_6 and R_7 have two common items namely, *Chocolate* and *Biscuit*. R_7 has a higher *w*-value. Support of R_7 (0.11) is much lower than that of R_6 (0.24). Hence R_7 is not able to account for the presence of {*Chocolate*,*Biscuit*} as much as R_6. Secondly, presence of *Milk* in R_7 further increases its *weakness*-value because R_7 is able to explain the presence of *Milk* in only 11 of the 50 transactions (22.0%) that contain *Milk*. However, a high support value does not guarantee a low *weakness*-value. R_3's *weakness*-value (Support=0.30,*w*=0.334146) demonstrates this. R_3's support though high is not sufficient to cover the presence of *Bread* and *Milk*.

Table 3 lists the clusters obtained through the average-linkage method [4]. Despite the difference (0.017523) in the *weakness*-values between R_{14} and R_5 being greater than the difference (0.010614) between R_{10} and R_6, the former pair merges earlier. R_{14} and R_5 being *weaker* rules leads to lesser inter-rule distance as compared to R_{10} and R_6.

A rule and its sub-rules being in different clusters may be due to the difference in support between a rule and its sub-rules. If the support values of a rule's items have wide variation, then different sub-rules may explain their constituents' presence to different extents. This difference in their *weakness*-values may place them in different clusters. Cluster configuration after Step 9 results in clusters C_{w1}:{R_{14},R_5,R_7} and C_{w2}:{R_{10},R_6} whose elements have an average *w*-values of 0.608383 and 0.141671 respectively. R_7 is a member of high-*weakness* C_{w1} while its sub-rules R_{14} and R_6 are members of clusters C_{w1} and low-weakness C_{w2} respectively. Support values of constituents *Milk* (0.50), *Chocolate* (0.33) and *Biscuit* (0.24) also show some variation. Thus, low-support coupled with high variation in the support values of its constituents might result in a *weak* rule.

Surprisingly, rules describing *Milk* (the most frequent item) belong to high-*weakness* clusters. None of the rules that contain *Milk* covers its presence to a substantial extent. High support of *Milk* also increases the *weakness* of low-support rules that contain it. Thus, a frequently occurring item may be present in many high-*weakness* rules if the item is purchased in many non-overlapping low-support market-baskets.

Another observation is with respect to rules in clusters that have relatively lower average *weakness*-values. Low-*weakness* clusters may not contain high-support rules. Consider C_{w2}:{R_{10},R_6}. Note that support of R_{10} (0.23) is quite close to support of its items *Pencil* (0.28) and *Eraser* (0.26). High explanatory power of such a rule is derived from its support value being close to the support values of its constituent items.

5. COMPARATIVE ANALYSIS AND DISCUSSION

Sahar [7] defines d_{SC}-distance on the basis of difference in rule's itemsets and overlap in the set of transactions that each rule covers. d_{SC} considers itemsets in antecedent/consequent in their entirety while d_w considers each item of a rule separately. Table 4 displays d_{SC}–based cluster configurations.

R_9 is a sub-rule of R_8 both having support 0.13. Their antecedents match completely. Hence contribution due to antecedent dissimilarity towards $d_{SC}(R_8,R_9)$ is 0. Also, R_9's consequent ({*Pencil*}) is a subset R_8's consequent ({*Pencil*,*Eraser*}). R_9 covers all transactions covered by R_8 thus increas-

Table 4. d_{SC} –based clustering

Step_No	Clusters
1	$\{\mathbf{R_9,R_8}\}$ [0.429]
2	$\{\mathbf{R_{12},R_2}\};\{R_9,R_8\}$ [0.437]
3	$\{\mathbf{R_5,R_1}\};\{R_{12},R_2\};\{R_9,R_8\}$ [0.442]
4	$\{\mathbf{R_{11},R_9,R_8}\};\{R_5,R_1\};\{R_2,R_{12}\}$ [1.098]
5	$\{\mathbf{R_4,R_5,R_1}\};\{R_{11},R_9,R_8\};\{R_2,R_{12}\}$ [1.892]
6	$\{\mathbf{R_{13},R_{12},R_2}\};\{R_4,R_5,R_1\};\{R_{11},R_9,R_8\}$ [1.958]
7	$\{\mathbf{R_{10},R_{11},R_9,R_8}\};\{R_{13},R_{12},R_2\};\{R_4,R_5,R_1\}$ [2.244]
8	$\{\mathbf{R_{14},R_6}\};\{R_{10},R_{11},R_9,R_8\};\{R_{13},R_{12},R_2\};\{R_4,R_5,R_1\}$ [2.313]
9	$\{\mathbf{R_{13},R_{12},R_3,R_2}\};\{R_{14},R_6\};\{R_{10},R_{11},R_9,R_8\};\{R_4,R_5,R_1\}$ [2.734]
10	$\{\mathbf{R_{13},R_{12},R_3,R_2,R_4,R_5,R_1}\};\{R_{14},R_6\};\{R_{10},R_{11},R_9,R_8\}$ [2.773]
11	$\{\mathbf{R_7,R_{14},R_6}\};\{R_{13},R_{12},R_3,R_2,R_4,R_5,R_1\};\{R_{10},R_{11},R_9,R_8\}$ [2.875]
12	$\{\mathbf{R_7,R_{14},R_6,R_{13},R_{12},R_3,R_2,R_4,R_5,R_1}\},\{R_{10},R_{11},R_9,R_8\}$ [3.980]
13	$\{\mathbf{R_7,R_{14},R_6,R_{13},R_{12},R_3,R_2,R_4,R_5,R_1,R_{10},R_{11},R_9,R_8}\}$ [4.437]

Note: Values in the brackets represent merging distance ing their similarity. Hence their low d_{SC}-value (0.429167). Hence R_8 and R_9 merge at Step 1.

d_{SC}-based clustering is useful in bringing together rules originating from the same portion of a database [7]. Here each cluster consists of rules whose items are members of the same or close domains. However, a rule and its sub-rules may vary a great deal on parameters like explanatory power, etc. Hence, a user may have to examine different clusters to find rules having the same specificity/generality.

Our scheme namely, groups rules having 'similar' values of *weakness* (similar explanatory power) irrespective of their domain. Characteristics like average-*weakness* may be used to define low-*weakness* clusters leading to appropriate clusters for further examination. Rules in other clusters need not be examined thus mitigating the rule immensity problem to some extent.

6. REFERENCES

1. Adomavicius, G., Tuzhilin, A.: Expert-Driven Validation of Rule-Based User Models in Personalization Applications. Data Mining and Knowledge Discovery. **5**, 1/2 (2001) 33-58.
2. Dong, G., Li, J.: Interestingness of Discovered Association Rules in Terms of Neighborhood-Based Unexpectedness. Proceedings of 2nd PAKDD, Springer-Verlag (1998) 72-86
3. Gupta, G. K., Strehl, A., Ghosh, J.: Distance-Based Clustering of Association Rules. Proceedings of Intelligent Engineering Systems through Artificial Neural Networks, (ANNIE 1999), ASME Press. Vol **9** (1999) 759-764
4. Jain, A. K., Murty, M.N., Flynn, P. J.: Data Clustering: A Review, ACM Computing Surveys. **31**, 3 (1999) 264-323
5. Jorge, A.: Hierarchical Clustering for thematic browsing and summarization of large sets of Association Rules, Proceedings of 2004 SIAM Conference on Data Mining (2004), http://www.siam.org/meetings/sdm04/proceedings/ sdm04_017.pdf
6. Lent, B., Swami, A., Widom, J.: Clustering Association Rules. Proceedings of Thirteenth International Conference on Data Engineering, Birmingham, UK. (April 1997) 220-231
7. Sahar, S.: Exploring Interestingness through Clustering: A Framework. Proceedings of IEEE International Conference on Data Mining (ICDM 2002), IEEE Computer Society Press. (2002) 677-680
8. Toivonen, H., Klemettinen, M., Ronkainen, P., Hatonen, K., Mannila, H.: Pruning and Grouping Discovered Association Rules. Proceedings of the MLnet Workshop on Statistics, Machine Learning and Knowledge Discovery in Databases, Herakhion, Crete, Greece, April, (1995)