Indirect Association : Mining Higher Order Dependencies in Data *

Pang-Ning Tan[1], Vipin Kumar[1], and Jaideep Srivastava[1]

Department of Computer Science,
University of Minnesota,
200 Union Street SE,
Minneapolis, MN 55455.
{ptan,kumar,srivasta}@cs.umn.edu

Abstract. This paper introduces a novel pattern called indirect association and examines its utility in various application domains. Existing algorithms for mining associations, such as Apriori, will only discover itemsets that have support above a user-defined threshold. Any itemsets with support below the minimum support requirement are filtered out. We believe that an infrequent pair of items can be useful if the items are related indirectly via some other set of items. In this paper, we propose an algorithm for deriving indirectly associated itempairs and demonstrate the potential application of these patterns in the retail, textual and stock market domains.

1 Introduction

In recent years, there has been considerable interest in extracting association rules from large databases [1]. Conceptually, an association rule indicates that the presence of a set of items in a transaction often implies the presence of other items in the same transaction. The problem of mining association rules is often decomposed into two subproblems : (1) discover all itemsets having support above a user-defined threshold, and (2) generate rules from these frequent itemsets. Under this formulation, any itemsets that fail the support threshold condition are considered to be uninteresting. However, we believe that some of the infrequent itemsets may provide useful insight about the data. Consider a pair of items, $\{a, b\}$, that seldom occurs together in the same transaction. If both items are *highly dependent* on the presence of another itemset, Y, then a and b are said to be indirectly associated via Y (Figure 1).

There are many potential applications for indirect associations. For market basket data, this method can be used to perform competitive analysis. For example, a and b may represent products of competing brands, such as Reebok and Nike. Suppose Reebok marketers are interested in expanding their current market share by attracting Nike customers through direct marketing campaigns. However, instead of promoting to every Nike customers, such a campaign can be

* supported by NSF ACI-9982274 and AHPCRC Contract No. DAAH04-95-C-0008.

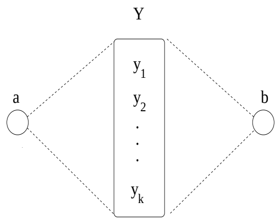

Fig. 1. Indirect Association between a and b via a mediating itemset Y.

made more effective, in terms of cost-benefit and lift analysis [6], by selecting a smaller target group whose buying behavior resemble that of Reebok customers. Indirect association provides an approach to characterize the group by identifying the set of items that are often bought by both groups of customers.

In the text domain, indirect association often corresponds to synonyms, antonyms or words that are used in the different contexts of another word. As an example, the words *coal* and *data* can be indirectly associated via *mining*. If a user queries on the word *mining*, the collection of documents returned often contains a mixture of both mining contexts. However, with indirect association, one can potentially identify explicitly the different ways in which the queried word appears in the corpus of text documents. Similarly, for stock market data, indirect association can help to identify the different set of events influencing the movement of a stock price.

The importance of indirect relationship between attributes of a dataset has been acknowledged by several authors [5, 4]. However, there has not been any direct attempts to explicitly derive such patterns. For example, in [5], Melamed observed that indirectly-associated words tend to reduce the accuracy of automated document translation systems by polluting the lexicon translation tables. Das et al. [4] introduced the notion of external similarity measure between attributes of a database relation. Essentially, external similarity is a measure of proximity between two attributes using the values in other columns, called the probe attributes. The notion of probe attribute is similar to our idea of mediator for indirect association. However, in [4], the role of a probe attribute is minimal; it is used only as far as determining the similarity between two attributes. On the other hand, a mediator is central to the concept of indirect association. Furthermore, probe attributes are chosen according to domain knowledge or constraints specified by a user [4]; whereas mediators are automatically derived from the observation data, as will be described in a later section.

2 Problem Formulation

Let $I = \{i_1, i_2, \cdots, i_d\}$ denotes a set of binary literals (called items) and T is the set of all transactions, $T = \{T_j \mid \forall j : T_j \subseteq I\}$. We will use upper case letters to represent itemsets (or sets of itemsets) and lower-case letters for individual items. Also, let $sup(X)$ denotes the support of an itemset, X.

Definition 1. *An itempair $\{a, b\}$ is indirectly associated via a mediator set Y if the following conditions hold :*

1. *$sup(a, b) < t_s$ (Itempair Support Condition)*
2. *There exists a non-empty set Y such that $\forall Y_i \in Y$:*
   (a) *$sup(a, Y_i) \geq t_f, sup(b, Y_i) \geq t_f$ (Mediator Support Condition).*
   (b) *$d(a, Y_i) \geq t_d,\ d(b, Y_i) \geq t_d$ where $d(p, Q)$ is a measure of the dependence between p and Q (Dependence Condition).*

Condition 1 is needed because an indirect association is significant only if both items rarely occur together in the same transaction. Otherwise, it makes more sense to characterize the pair in terms of their direct association. An alternative to this condition is to test for independence between the two items. However, it is often the case that independent or negatively correlated itempairs tend to have low support values. Therefore, Condition 1 will effectively consider only itempairs that are slightly or negatively correlated.

Condition 2(a) can be used to guarantee the statistical significance of the mediator set. In particular, for market basket data, the support of an itemset justifies the feasibility of promoting the items together. Support also has a nice downward closure property which allows us to prune the combinatorial search space of the problem.

Condition 2(b) ensures that only items that are highly dependent on both a and b are used to form the mediator set. Over the years, many measures have been proposed to represent the degree of dependence between attributes of a dataset. One such measure is Pearson's linear correlation coefficient, ϕ. For binary variables, it can be shown that within certain range of support values [1], the correlation coefficient $\phi_{x,y}$ can be expressed in terms of the interest factor [3, 2], $I(x, y) \equiv \frac{P(x,y)}{P(x)P(y)}$, and support, i.e. :

$$\phi_{x,y} \approx \sqrt{\mathrm{I(x, y) \times sup(x, y)}} \ .$$

We will use the right-hand side of this expression, called the IS measure, as the dependence measure in Condition 2(b). This measure is desirable because it takes into account both the interestingness and support aspects of a pattern. However, our general framework can accommodate other measures, such as Piatetsky-Shapiro's rule-interest, J-measure and Gini index, which have been shown to be equally good at capturing statistical correlation [7].

3 Algorithm

An algorithm for mining indirect association is given in Table 1. Initially, an itempair support matrix S is constructed by scanning the entire database (step 2). Next, S will be used to prune the itempair space (step 3) based on the following criteria : (1) If the support of a is below t_f or a does not belong to any

[1] when $P(x) \ll 1$, $P(y) \ll 1$ and $\frac{P(x,y)}{P(x)P(y)} \gg 1$.

frequent itempairs, then the mediator set for a will always be empty. (2) Any itempairs that violate Condition 1 will be removed. There are two main phases in the FindMediator step : candidate generation and pruning of the mediator. Basically, it is assumed that a lattice of frequent itemsets, FI, has been generated using standard algorithm such as Apriori. During candidate generation, it will find all candidate mediators, $Y_i \subseteq I - \{a, b\}$, such that $\{a\} \cup Y_i \in FI$ and $\{b\} \cup Y_i \in FI$. The IS measure for each Y_i is then computed. Finally, during the pruning phase, candidates that fail Condition 2(b) are removed.

4 Experimental Results

To demonstrate the utility of indirect associations, experiments were carried out using datasets from three application domains : text, retail and stock market data. Table 2 shows a summary of the datasets used along with the thresholds chosen for our experiments.

Reuters-21578 Distribution 1.0 newswire articles.

This dataset contains a collection of financial and commodity news articles that appeared on Reuters newswire in 1987.[2] Articles from both categories are preprocessed by removing stopwords and stemming each word to its root form. Table 3 shows some of the indirectly associated stemmed words from both collections of news articles. Most of the indirect associations represent the different contexts in which a term may appear in the news collection. For example, the indirect association between Soviet and worker refers to two separate news threads about union : one corresponds to news stories about Soviet Union while the other involves articles on labor unions.

Retail data

The retail data was obtained courtesy of Fingerhut Corp. As expected, most of the indirect associations correspond to pairs of competing items (Table 4). Instead of doing competitive analysis, we are interested in discovering *surprising* patterns. A pattern is surprising if items belonging to competing product categories are directly associated. Hence, the first itempair is not surprising because it relates two products of different sizes via items that do not have a size distinction. Nevertheless, this type of pattern can be useful to determine what products should be bundled together for upsale promotions. The second itempair involves two products with competing design logos (checkered flag versus Nascar drivers). Since each logo has its own matching comforter, sheet, pillow case, etc., it is not surprising that their joint support is low. However, unlike the previous example, this pattern is unexpected because we do not expect checkered flag comforters and Nascar drivers wallpapers (the mediator) to have a large support value. Upon closer examination, we found that the reason their observed support is high is because the product catalog does not offer any checkered flag

[2] available at http://www.research.att.com/˜lewis.

Table 1. Basic algorithm for mining indirect association between itempairs.

1. let $S = [sup(a, b)]$ denotes the support matrix for all itempairs (a, b).
2. For each transaction $t_i \in T$, UpdateSupportMatrix(t_i, S).
3. prune the itempair space.
4. for each remaining itempair (a, b) :
   4a. $Y \leftarrow$ FindMediator(S, a, b, t_d, t_f)
   4b. if $Y = \emptyset$, go to step 4 else $\{a, b\}$ is indirectly associated via Y.

Table 2. Summary of dataset parameters and results.

Dataset	t_s	t_f	t_d	n	|T|	# Freq pairs	Indirect pairs
Reuters(Finance)	0.15%	1.5%	0.3	2886	2005	10877	99
Reuters(Commodity)	0.15%	1.5%	0.3	3785	2308	6621	34
Retail Data	0.01%	0.1%	0.1	14462	58565	1174	59
S&P-500	0.25%	2.5%	0.2	976	716	13229	262

Table 3. Indirect association for Reuters-21578 Finance (left) and Commodity (right) datasets.

a	b	Y_i	$d(a, Y_i)$	$d(b, Y_i)$
suppli	shortag	monei	0.4940	0.4597
partner	temporari	trad	0.3932	0.3043
partner	dealer	trad	0.3932	0.3064
shortag	growth	forecast	0.3429	0.3502
shortag	inflat	forecast	0.3429	0.3241

a	b	Y_i	$d(a, Y_i)$	$d(b, Y_i)$
soviet	strike	union	0.6248	0.4149
soviet	worker	union	0.6248	0.3615
opec	ico	quota	0.4112	0.5539
opec	coffee	quota	0.4112	0.4515
iran	tax	oil	0.3225	0.3312

Table 4. Indirect Association for retail data

a	b	Y_i	$d(a, Y_i)$	$d(b, Y_i)$
Comforter Queen (Madrid)	Comforter King (Madrid)	Drapes (Madrid)	0.5634	0.4698
		Valance (Madrid)	0.5771	0.4708
		Pillow (Madrid)	0.5840	0.4446
Comforter Twin (Checkered Flag)	Sheets Twin (Nascar Drivers)	Border Wallpaper (Nascar Drivers)	0.3322	0.2517
Comforter Twin (Checkered Flag)	Curtains (Nascar Drivers)	Border Wallpaper (Nascar Drivers)	0.3322	0.2481
Playstn W/Crash2	Playstn Controller	Playstn memory card	0.2784	0.4448

Table 5. Indirect Association for S&P 500 data

a	b	Y_i	$d(a, Y_i)$	$d(b, Y_i)$
ibm-up	yell-up	lsi-up	0.3244	0.2200
		mu-up	0.2507	0.2005
hwp-down	txn-up	gnt-up	0.2002	0.2290
amgn-down	gt-down	digi-down	0.2303	0.2313
		s-down	0.2548	0.2826
oxy-down	ph-down	adsk-down	0.2740	0.2334
axp-down	nke-up	lsi-down	0.2197	0.2093

wallpapers. As a result, many customers who buy checkered flag comforters end up buying Nascar drivers wallpapers.

S&P 500 stock market data.

The dataset represents the daily fluctuation of share prices for S&P-500 stocks from Jan. 1994 to Oct. 1996. Each stock is represented by two attributes, X-up and X-down. The value of X-up (or down) is 1 if the closing price for stock X is significantly higher (lower), by at least 2%, than its previous closing price. Indirect associations can be used for event segmentation, where one can determine the set of events that are causing the price of a stock to move up or down. For example, the first indirect association in Table 5 relates IBM-up with YELL-up via LSI-up and MU-up. IBM is a company that provides various customer solution in information technology while YELL (Yellow Corp) is involved in the transportation business. The mediator contains the stocks of two semiconductor companies, LSI Logic and Micron Technology. This pattern indicates that events involving LSI-up and MU-up, can be partitioned into three disjoint sets - one involving IBM-up, another associated which YELL-up; and a third set of events not related to IBM-up nor YELL-up.

5 Conclusions

In summary, the above results show that indirect association can provide meaningful insight into the data. Such knowledge can not be derived from association rules alone because it involves infrequent itempairs and require analysis of higher order dependencies between items. Due to space limitation, details regarding the complexity analysis of our algorithm and threshold selection issues have been omitted. Interested readers can read the expanded version of this paper in [8].

References

1. R. Agrawal, T. Imielinski, and A. Swami. Database mining: a performance perspective. *IEEE Transactions on Knowledge and Data Engineering*, 5:914–925, 1993.
2. T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets. Using association rules for product assortment decisions : A case study. In *Proc. KDD'99*, San Diego, August 1999.
3. S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proc. SIGMOD'97*, Tucson, AZ, 1997.
4. G. Das, H. Mannila, and P. Ronkainen. Similarity of attributes by external probes. In *Proc. KDD'98*, New York, NY, 1998.
5. D. Melamed. Automatic construction of clean broad-coverage translation lexicons. In *2nd Conf. of the Association for Machine Translation in the Americas*, 1996.
6. G. Piatetsky-Shapiro and B. Masand. Estimating campaign benefits and modeling lift. In *Proc. KDD'99*, San Diego, 1999.
7. P.N. Tan and V. Kumar. Interestingness measures for association patterns : A perspective. Technical Report TR00-036, University of Minnesota, 2000.
8. P.N. Tan, V. Kumar, and J. Srivastava. Indirect association : Mining higher order dependencies in data. Technical Report TR00-037, University of Minnesota, 2000.