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 the concept of indirect association between items and examines its utility in various application domains. Existing algorithms for mining association rules, such as Apriori, will only discover itemsets that have support above a user-defined threshold. Any itemsets that fall below the minimum support requirement are filtered out. We will show that some of the removed itemsets may provide useful insight into the data. Consider a pair of items (a, b) with a low support value. If there is an itemset Y such that the presence of a and b are highly dependent on items in Y, then (a, b) are said to be indirectly associated via Y. We have identified many potential applications for indirect associations. In market basket scenario, these patterns can be used to perform competitive analysis among products. For text documents, indirect associations can be used to identify synonyms, antonyms or words that appear in the different contexts of another word. We will present a formal framework for describing indirect association and propose an algorithm for mining such patterns. Finally, we will demonstrate the benefits of mining these patterns based on empirical results obtained from retail, textual and stock market data.

1 Introduction

In recent years, there has been considerable interest in extracting association rules [AIS93b,AIS93a] from large databases. Conceptually, an association rule indicates that the presence of a set of items (called itemset) in a transaction often implies the presence of other items in the same transaction. Current research effort has focussed on developing more efficient algorithms for mining these patterns [AS94,PCY95,SON95] as well as extending the concept beyond the single transaction, binary literal paradigm [TLHF99,FLYH99].

The problem of mining association rules is often decomposed into two separate tasks : (1) discover all itemsets having support [1] above a user-defined threshold, and (2) generate rules from these *f*requent 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 co-occurs together in the same transaction. If both items are *h*ighly dependent on the presence of another itemset, Y, then the pair (a, b) is said to be indirectly associated via Y (Figure 1). In this paper, we will investigate the problem of mining indirect association and examine its utility in various application domains. Note that unlike typical association rules, an indirect association pattern is characterized by the two items participating in this relationship, along with all other items mediating the interaction.

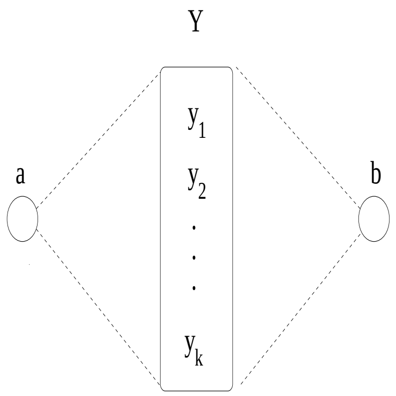

Fig. 1. Indirect Association between a and b via a mediating itemset Y.

We motivate the usefulness of indirect association with several examples. For market basket data, this method can be used to perform competitive analysis of products. As an example, a and b may represent competing brands of a footwear

[1] Support of an itemset X denotes the fraction of transactions containing the itemset X.

product, such as Reebok and Nike. [2] Suppose Reebok marketing executives are interested in expanding their current market share by attracting customers of their competitor through direct marketing campaigns. However, instead of promoting to every Nike customers, such a campaign can be made more effective, in terms of cost-benefit and lift analysis [LL98,PSM99], by selecting a smaller target group whose buying behavior strongly resemble that of Reebok customers. This is where indirect association can help, by identifying what are the items that are highly dependent and frequently bought together with Nike and Reebok products.

For text documents, indirect association between a pair of words often correspond to synonyms, antonyms or words that are present in the different contexts of another word. For 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. Explicit discovery of indirect association provides an opportunity to segment the query results into various contexts of the queried term.

Similarly, for stock market data, indirect association can be used to identify the different sets of events influencing the upward or downward movement of a stock price. For example, Redhat-Up and Intel-Down may be indirectly associated via Microsoft-Down. One can interpret this pattern as saying that the event Microsoft-Down is associated with at least two other disjoint sets of events : one which is causing the stock price of its competitor to go up while the other may correspond to the decline of other technology group stocks.

This paper is organized in the following way. In Section 2, a formal definition of indirect association is presented. Next, an algorithm for mining such pattern is given, followed by empirical results showing its utility in various application domains. Finally, Section 5 concludes with a summary of our work and suggestions for future research.

2 Indirect Association

2.1 Related Work

The importance of indirect association between attributes of a dataset has been recognized by several authors [Mel96,DMR98]. However, to our knowledge, there has not been any direct attempts to explicitly derive such patterns from large datasets.

[Mel96] observed that automated document translation systems tend to produce lexicon translation tables that are full of indirectly-associated words. A lexicon translation table encodes the probability that two words from different languages being semantically equivalent to one another. The presence of indirect association can pollute the resulting tables, thereby reducing the overall precision of the system. An iterative strategy was proposed in [Mel96] to clean up

[2] The use of these brand names are for illustrative purpose only.

existing translation tables by finding only the most probable translations for a given word.

[DMR98] introduced the notion of internal and external measures of similarity between attributes of a database relation. Internal similarity between two attributes a and b is a measure whose value depends only on the values of a and b columns. Conversely, external measure takes into account data from other columns (called the probe attributes). Their notion of probe attributes is similar to our idea of mediators for indirect association. However, their sole purpose of using probe attributes is to perform attribute clustering.

Indirect association is closely related to the notion of negative association rules [SON98]. In both cases, we are dealing with itemsets that do not have sufficiently high support. A negative association rule discovers what are the items a customer will not likely buy given that he/she buys a certain set of other items. Typically, the number of negative association rules can be prohibitively large and the majority of them are not interesting to a data analyst. [SON98] proposed the use of domain knowledge, in the form of item taxonomy, to decide what constitutes an interesting negative association rule. The intuition here is that items belonging to the same parent node in a taxonomy are expected to have similar types of associations with other items. If the observed support is significantly smaller than its expected value, then we conclude that a negative association exists between the items. Again, unlike indirect association, these type of regularities do not specifically look for mediating elements.

Another related area is the study of functional dependencies in relational databases. Functional dependencies are relationships that exist between attributes of a relation. However, the emphasis of functional dependencies is to find dependent and independent attributes for applications such as semantic query optimization [Wed92] and reverse engineering [TBSH93].

2.2 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, $T = \{t_j \mid \forall j : t_j \subseteq I\}$. Below, we give a formal definition of indirect association.

Definition 1. *An itempair $\{a, b\}$ is indirectly associated via a mediator 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).*

The above thresholds are called itempair support threshold (t_s), dependence threshold (t_d) and frequent itemset threshold (t_f) respectively. At first glance, it may seem that there are too many parameters needed in the above formulation. In practice, we only need two threshold conditions, t_f and t_d (similar to support and confidence thresholds for association rules). t_s can always be chosen to be some reasonable constant fraction of t_f. [3]

Condition 1 is needed because an indirect relationship between two items 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 two items. Nevertheless, it is often the case that itempairs consisting of independent (or negatively correlated) items tend to have low support values in many natural datasets.

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 affects the amount of revenue generated and justifies the feasibility of a marketing decision [BSVW99]. Moreover, support has a nice downward closure property which allow us to prune the combinatorial search space of the problem.

Condition 2(b) ensures that only items that are highly dependent on the presence of a and b will be used to form the mediator set. Over the years, many measures have been proposed to quantify the degree of dependence between attributes of a dataset. From statistics, the χ^2 test is often used for this purpose.. However, the drawback of this approach is that it does not measure the strength of dependencies between items [BMS97,WH75]. Furthermore, the χ^2 statistic depends on the number of transactions in the database. As a result, other statistical measures of association are often used, including Pearson's ϕ coefficient, Goodman and Kruskal's λ, Yule's Q and Y coefficients, etc [Rey77].

In the original formulation of association rule mining, confidence was chosen to measure the goodness of a rule. However, [BMS97] showed that this measure may produce counter-intuitive results especially when an itemset exhibits a strong negative correlation among its items. Interest factor is another measure that has been used quite extensively to quantify the strength of dependency among items [BMS97,BSVW99,CC99].

Definition 2. *The interest factor between two items X and Y is :*

$$I(X,Y) \equiv \frac{P(X,Y)}{P(X)P(Y)} \tag{1}$$

Even though interest factor is an appropriate measure from a statistical viewpoint, it may not be interesting within the microeconomic framework of a retailer [BSVW99].

Other objective measures of interestingness from data mining and machine learning literature include Piatetsky-Shapiro's rule-interest, Gini index, conviction, laplace, J-measure, etc. An important criteria for a good objective measure lies in its ability to capture the statistical notion of correlation. For binary variables, correlation coefficient is given by :

$$\rho_{X,Y} = \frac{E(X,Y) - E(X)E(Y)}{\sqrt{E(X)\ (1-E(X))}\sqrt{E(Y)\ (1-E(Y))}} \ . \tag{2}$$

[3] In our experiments, we chose $t_s = t_f/10$.

where $E(\cdot)$ is the expected value of its argument. It can be shown that for certain range of support values [4],

$$\rho_{X,Y} \approx \sqrt{\mathrm{I(x,y)} \times \mathrm{sup(x,y)}/|\mathrm{T}|}\ .$$

We will denote the right-hand side of this expression as the IS measure.

Figure 2 shows the relationship between $\rho_{X,Y}$ and I factor for various datasets used in our experiments. The graph was obtained by comparing the correlation coefficient against I factor for all itempairs that pass the minimum support threshold given in Table 2. Figure 3 is the corresponding graph for $\rho_{X,Y}$ and IS measure. Note the high linearity of these graphs, particularly for retail data. This suggests that IS measure is indeed a reasonable estimate of statistical correlation. Moreover, it is a desirable metric since it takes into account both the interestingness and support aspects of a pattern. In this paper, we will use IS as the dependence measure for Condition 2(b).

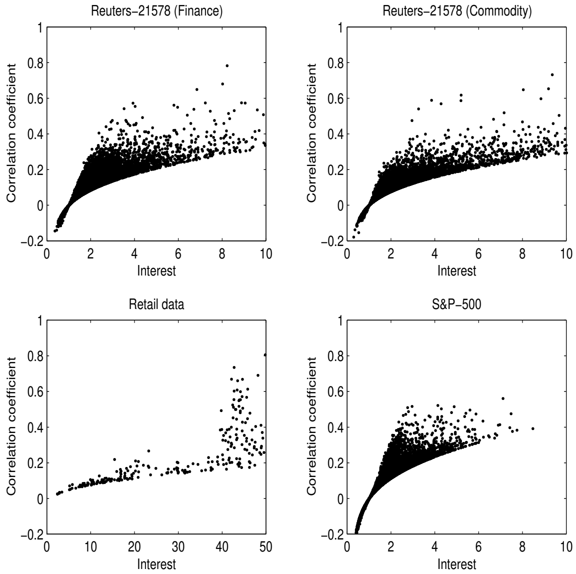

Fig. 2. Comparison between I factor and $\rho_{X,Y}$ for frequent itempairs in various datasets.

Mediator of Indirect Association. Another issue we need to address is what constitutes an element of the mediator set. A naive approach is to treat single items as mediating elements. For example, in the market basket scenario, if

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

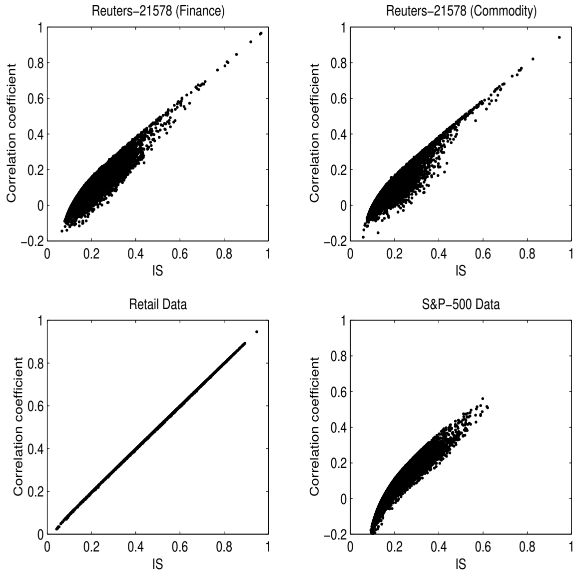

Fig. 3. Comparison between IS and $\rho_{X,Y}$ for frequent itempairs in various datasets.

the purchase of coffee and tea is highly dependent on the purchase of sugar and cream, then $Y = \{\{\text{sugar}\}, \{\text{cream}\}\}$ may serve as the mediator between coffee and tea. One can extend this definition to include itemsets of larger size. For example, in the text domain, individual words may not have significant dependencies compare to phrases of words. Other forms of representation are also possible. For example, a mediating element can contain both positive and negative literals $(\bar{p}, q, \bar{r})$ as well as disjunctive items $(p \vee q \vee r)$. In this paper, we will focus on conjunction of positive literals as mediating elements.

Indirect Affinity. It is often useful to quantify the overall affinity between two indirectly-associated items. Such a metric can be used to rank the discovered patterns according to the strength of their indirect association. If the mediator has only one element y, then, the overall affinity between a and b is :

$$Aff(a, b|y) = \min(IS(a, y), IS(b, y)) \tag{3}$$

where $d(x, y)$ is the dependence measure between x and y. [5].

For mediator set with more than one element, there are several ways to define the overall affinity between a and b. In this paper, we will use the lower bound of this value, i.e. :

$$Aff(a, b|Y) \geq \max_{y_i \in Y}(\min(f(a, y_i), f(b, y_i))) \tag{4}$$

[5] Other functions based on arithmetic or geometric means can also be used.

3 Algorithm

3.1 Main Algorithm

An algorithm for mining indirect association between itempairs is given in Table 1. Initially, an itempair support matrix S is constructed by scanning the entire database (step 2). The sum of each row in S will give the support count of each individual item. Next, the support matrix will be used to prune the itempair space (step 3) based on the following criteria :

- If the support of a is below t_f, it will not appear in any frequent itemsets. Therefore, we can remove all itempairs involving a from our consideration.
- If the support of a is above t_f, but a does not belong to any frequent itempairs, then the mediator set for a will always be empty. Again, we can ignore any itempairs involving a from further consideration.
- Any itempairs that violate the itempair support condition can be removed.

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

Indirect Association Algorithm :

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 next itempair.
   4c. else $IP \leftarrow (a, b|Y)$
5. Filtering or ranking itempairs.

The key step of this algorithm lies in step 4a. There are two major phases in this step : candidate generation and pruning of the mediator. Basically, it assumes that a lattice of frequent itemsets, FI, has been generated using standard algorithm such as Apriori. During each pass of candidate generation, it will find all frequent itemsets, $y_i \subseteq I - \{a, b\}$, such that both $\{a\} \cup y_i \in FI$ and $\{b\} \cup y_i \in FI$. Then, it will compute the IS measure for each y_i with respect to both a and b. Finally, during the pruning step, all candidate itemsets that fail the mediator dependence condition will be removed. If the resulting mediator set is empty, then we conclude that the itempair is not indirectly associated (step 4b). Otherwise, we will insert the pair along with its mediator into IP. The indirect itempairs can then be ranked according to their IS measure or filtered using other subjective measures (step 5).

3.2 Complexity Analysis

We will now briefly discuss the complexity of our algorithm. The UpdateSupportMatrix function requires a single scan over the entire database. Assuming that the matrix update operation takes O(1) and the maximum width of a transaction is k, this operation takes at most $\mathrm{O}(k^2|T|)$. The pruning step can be executed in $\mathrm{O}(n^2)$ where n is the total number of items. Next, for each infrequent itempair (a, b), the most expensive operation in the FindMediator function, T_m is the candidate generation. The complexity of this step depends on the number of frequent itemsets containing a or b. There is also an overall cost of computing the frequent itemset lattice, T_c.

The ranking step requires us to first compute the indirect affinity between every discovered itempairs. Suppose w is the maximum number of mediating itemsets between a pair of indirectly associated items. In the worst-case scenario, this step will take $\mathrm{O}(n^2 w)$ plus an additional $\mathrm{O}(n^2)$ to sort the itempairs. Hence, the overall worst-case complexity of this algorithm is $\mathrm{O}(k^2|T| + wn^2 + n^2 T_m + T_c)$.

3.3 Alternative scheme for finding single item mediators

An alternative way for finding indirectly associated itempairs with single item mediators is by using a bottom-up approach. Essentially, the algorithm assumes that all items can be potential mediators. For each such item, y, the goal is to find all neighboring items x_i, such that (x_i, y) satisfies both mediator support and dependence conditions. Then, for every pair of such neighbors (x_i, x_j), we will declare them as indirectly associated if the support of these itempairs are below t_s. This approach is useful for finding indirect association in text and stock datasets, where the goal is to find the different context (or events) associated with the mediating element.

4 Experimental Results

To demonstrate the utility of indirect associations, experiments were carried out using datasets from three real-world domains : text, retail and stock data. Table 2 shows the parameters of each dataset along with the threshold values used. Figure 4 illustrates the effect of using various t_f values on the number of frequent and indirectly associated itempairs generated by our algorithm. These experiments were performed on a 500-MHz Pentium III machine with 256MB main memory.

4.1 Reuters-21578 Distribution 1.0 newswire articles.

This dataset consists of a collection of news articles that appeared on Reuters newswire in 1987.[6] Initially, two distinct subsets of the dataset were chosen : financial articles and commodity articles. The finance category contains articles

[6] available at http://www.research.att.com/~lewis

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

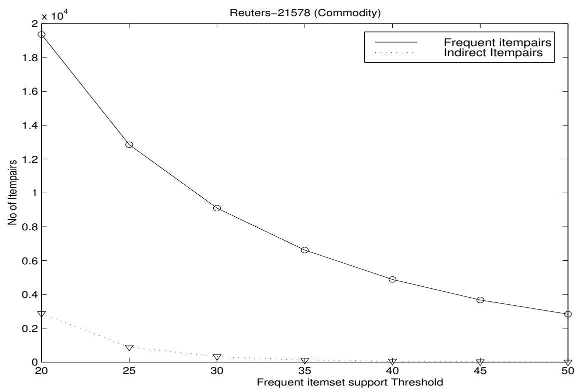

Fig. 4. Number of frequent and indirect itempairs versus frequent itemset thresholds, t_f

on topics such as trade, retail, money, interest, etc., whereas the commodity category encompasses subject areas such as rubber, sugar, gold, coffee and gas. Articles in both categories are preprocessed by removing stopwords and stemming each word to its root form.

The finance category contains 546 unique, stemmed words from 2005 news articles. Using our proposed algorithm, 99 indirectly associated itempairs are discovered. Table 3 shows some of the discovered itempairs, ranked according to their IS measure. As expected, most of the word pairs represent the different contexts of a word. For example, the word forecast often appears in two different contexts : forecast of product shortage and forecast of growth in the economy. Other examples include temporary agreement versus negotiate for agreement, trade dealer versus trade partner and supply of money versus shortage of money.

Table 3. Indirect Association between stemmed word pairs for Reuters-21578 Finance dataset.

#	a	b	y	$d(a,y)$	$d(b,y)$	IS rank
1	suppli	shortag	monei	0.4940	0.4597	2
2	partner	temporari	trad	0.3932	0.3043	8
3	partner	dealer	trad	0.3932	0.3064	77
4	shortag	growth	forecast	0.3429	0.3502	27
5	shortag	inflat	forecast	0.3429	0.3241	50
6	shortag	econom	forecast	0.3429	0.3167	61
7	negoti	temporari	agreem	0.3625	0.3337	37

There are 992 stemmed words and 2308 articles in the Reuters commodity news dataset. We found 34 indirectly associated word pairs satisfying the threshold conditions (Table 4). Among the interesting pairs include Soviet Union versus worker's union, Soviet Union versus union strike, and OPEC [7] quota versus ICO [8] quota. Co-incidently, using WordNet 1.6, we found two of the distinct meanings for union are "*of trade unions*" and "*a political unit formed from previously independent people or organizations; the Soviet Union*".

4.2 Retail data

The retail data was obtained courtesy of Fingerhut Corp. Table 5 illustrates some of the indirect associations discovered from this dataset. As expected, most of them correspond to pairs of competing items. This provides a great opportunity to do competitive analysis as described in Section 1. A second type of analysis is to discover *surprising* patterns. The intuition here is that items that belong to the same item category are expected to have similar types of association with other items [SON98].

[7] Organization of Petroleum Exporting Countries
[8] International Coffee Organization

Table 4. Indirect Association between stemmed word pairs for Reuters-21578 Commodity dataset.

#	a	b	y	$d(a, y)$	$d(b, y)$	IS rank
1	corpor	farmer	agricultur program	0.4946	0.4651	
			agricultur dlr	0.5160	0.4901	1
2	soviet	strike	union	0.6248	0.4149	9
3	soviet	worker	union	0.6248	0.3615	17
4	opec	ico	quota	0.4112	0.5539	10
5	opec	coffee	quota	0.4112	0.4515	11
6	iran	tax	oil	0.3225	0.3312	24
7	germani	texa	west	0.5173	0.3197	26

Table 5. Indirect Association for retail data

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

The first itempair relates two products of different sizes (King versus Queen size). The mediator set contains related products from the same manufacturer, but without any size distinction. Itempairs such as this are often uninteresting because they are not actionable nor do they reveal any surprising knowledge.

The second itempair involves two products with different design logos (checkered flag versus Nascar drivers). Each logo has its own matching comforter, sheet, pillow case, etc. Hence, it is not surprising that the support value of this pair is low. However, unlike the previous example, the mediator set is made up of wallpapers that belong to one of the two product groups. Such a pattern is surprising because we do not expect checkered flag comforters and Nascar drivers wallpapers (the mediator) to have a large support value. Therefore, this pattern is potentially interesting to a data analyst. Upon closer examination, we found that the reason their observed support is high is because the product catalog does not offer any checkered flag wallpapers. As a result, most of the customers who buy checkered flag comforters end up buying Nascar drivers wallpapers.

The third itempair is also interesting for the same reason as before. Furthermore, this pattern is potentially actionable since the product catalog does not offer any checkered flag curtains. The marketing executive will be interested in knowing why customers who buy checkered flag products are not buying Nascar-related curtains. Is it due to a lack of consumer awareness or simply because Nascar-related curtains are not appealing to checkered-flag customers. If the former is true, this provides an opportunity for cross-selling. On the other hand, if the latter is the case, such information will be a useful feedback to the manufacturer of Nascar products.

Summarizing, the above examples show that indirect association can provide meaningful insight into market basket data. Such knowledge can not be derived from association rules alone because they require analysis of higher order dependencies and they are derived from infrequent itempairs.

4.3 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. Some of the indirect associations found for this dataset are shown in Table 6. Indirect associations can be used for event partitioning, 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 relates the event IBM-up with YELL-up via LSI-up and MU-up. IBM (International Business Machines) 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 (MU). 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.

Table 6. Indirect Association between itempairs for S&P 500 data

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

5 Conclusions

This paper proposes a framework for describing indirect association between items in a database of transactions. An itempair is said to be indirectly associated if the support of the pair is low, yet, there exists a mediator set such that both items are highly dependent on the mediating elements. We showed that our measure of dependence is a good way to estimate the correlation between frequently occuring items. We have also provided a simple way to express the overall affinity between these itempairs. An algorithm for mining these patterns is also presented. Finally, we showed how this novel idea is applicable in various real-world applications.

For future research, there are several unresolved issues we need to address. First of all, it is possible to extend our work to discover indirectly associated itemsets rather than between a pair of items. In Section4, we have suggested one such approach by combining the itempairs found into larger itemsets. We hope to explore other possible ways to do this. Further studies are also needed to define a more exact measure of indirect affinity between itemsets with multiple mediating elements. The discontinuous nature of the minimum and maximum functions make it difficult to combine the itemsets in a natural way. Also, our current implementation assumes that the itemset lattice can be stored in the main memory. Therefore, scalability issue is a major concern. Threshold selection is another issue that needs further investigation.

6 Acknowledgement

We are grateful to Fingerhut Corp, in particular, Deb Campbell and Nadav Cassuto, for providing us with the retail dataset. We would also like to thank Sam Han for providing us with pre-processed version of the text and stock market data; George Karypis and Mahesh Joshi for many wonderful discussions.

References

[AIS93a] R. Agrawal, T. Imielinski, and A. Swami. Database mining: a performance perspective. *IEEE Transactions on Knowledge and Data Engineering*, 5:914–925, 1993.

[AIS93b] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. ACM SIGMOD Intl. Conf. Management of Data*, pages 207–216, Washington D.C., USA, 1993.

[AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, pages 487–499, Santiago, Chile, 1994.

[BMS97] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proc. ACM SIGMOD Intl. Conf. Management of Data*, pages 265–276, Tucson, AZ, 1997.

[BSVW99] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets. Using association rules for product assortment decisions : A case study. In *Proc. of the Fifth ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining*, pages 254–260, San Diego, Calif, August 1999.

[CC99] Robert Cooley Chris Clifton. Topcat: Data mining for topic identification in a text corpus. In *Proceedings of the 3rd European Conference of Principles and Practice of Knowledge Discovery in Databases*, 1999.

[DMR98] G. Das, H. Mannila, and P. Ronkainen. Similarity of attributes by external probes. In *Proc. of the Fourth ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining*, pages 23–29, New York, NY, 1998.

[FLYH99] L. Feng, H.J. Lu, J.X. Yu, and J. Han. Mining inter-transaction associations with templates. In *Proc. 1999 Int. Conf. on Information and Knowledge Management (CIKM'99)*, pages 225–233, Kansas City, Missouri, Nov 1999.

[KMR+94] M. Klemettinen, H. Mannila, P. Ronkainen, T. Toivonen, and A. Verkamo. Finding interesting rules from large sets of discovered association rules. In *Proc. 3rd Int. Conf. Information and Knowledge Management*, pages 401–408, Gaithersburg, Maryland, Nov 1994.

[LL98] C.X. Ling and C. Li. Data mining for direct marketing: Problems and solutions. In *Proc. of the Fourth ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining*, pages 73–79, New York, NY, 1998.

[Mel96] D. Melamed. Automatic construction of clean broad-coverage translation lexicons. In *2nd Conference of the Association for Machine Translation in the Americas (ATMA 96)*, 1996.

[PCY95] J.S. Park, M.S. Chen, and P.S. Yu. An effective hash-based algorithm for mining association rules. *SIGMOD Record*, 25(2):175–186, 1995.

[PSM99] G. Piatetsky-Shapiro and B. Masand. Estimating campaign benefits and modeling lift. In *Proc. of the Fifth ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining*, pages 185–193, San Diego, CA, 1999.

[Rey77] H.T. Reynolds. *The Analysis of Cross-Classifications*. Macmillan Publishing Co., New York, 1977.

[SON95] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. In *Proc. of the 21st Int. Conf. on Very Large Databases (VLDB'95)*, Zurich, Switzerland, Sept 1995.

[SON98] A. Savasere, E. Omiecinski, and S. Navathe. Mining for strong negative associations in a large database of customer transactions. In *Proc. of the 14th International Conference on Data Engineering*, pages 494–502, Orlando, Florida, February 1998.

[TBSH93] Z. Tari, O. Bukhres, J. Stokes, and S. Hammoudi. The reengineering of relational databases based on key and data correlations. In S. Spaccapietra and F. Maryanski, editors, *Searching for Semantics: Data Mining, Reverse Engineering, etc.* Chapman and Hall, 1993.

[TLHF99] Anthony Tung, Hongjun Lu, Jiawei Han, and Ling Feng. Breaking the barrier of transactions: Mining inter-transaction association rules. In *Proc. of the Fifth Int'l Conference on Knowledge Discovery in Databases and Data Mining*, pages 297–301, San Diego, CA, August 1999.

[Wed92] G.E. Weddell. Reasoning about functional dependencies generalized for semantic data models. *ACM Transactions on Database Systems*, 17(1):32–64, March 1992.

[WH75] R. Winkler and W. Hays. *Statistics: Probability, Inference and Decision*. Holt, Rinehart & Winston, New York, second edition, 1975.