Market-Basket Analysis with Principal Component Analysis: An Exploration

Kitty S.Y. Chiu, Robert W.P. Luk, Keith C.C. Chan and Korris F.L. Chung

Department of Computing, Hong Kong Polytechnic University

Hong Kong, China

{cssychiu, csrluk, cskcchan, cskchung}@comp.polyu.edu.hk

Abstract-- **Market-basket analysis is a well-known business problem, which can be (partially) solved computationally using association rules, mined from transaction data to maximize cross-selling effects. Here, we model the market-basket analysis as a finite mixture density of human consumption behavior according to social and cultural events. This leads to the use of principle component analysis and possibly mixture density analysis of transaction data, which was not apparent before. We compare PCA and association rules mined from a set of benchmark transaction data, to explore similarities and differences between these two data exploration tools.**

Keywords: **market-basket, association rules, PCA, mixture density**

I. INTRODUCTION

The market-basket problem is a well known business problem in which consumers buying behavior is influenced by alternative products and related products. With the availability of accurate and efficient capture of transaction data, computational analysis of transaction data can discover interesting consumer spending patterns. Mining association rules [1,2] is a well known and important tool to discover dependencies in product sales. By discovering these dependencies, it is possible to maximize the cross-selling effect [3] of related products, increasing transaction volume and therefore increasing total profit. The appropriate use of cross-selling capabilities has implications in other business activities, like inventory/warehouse management, enhancing user satisfaction, etc. Hence, it is important to discover product dependencies [4]. While association rules have been quite successful in solving the market-basket problem and discovering dependencies, higher-order dependencies are hard to find because of the nature of the discovery algorithm (e.g. Apriori), in which successive higher order association rules are pruned [5] due to computational cost and the lack of evidence to support such higher order association rules.

Principal component analysis (PCA) and the more general techniques of finding mixture densities are well known multivariate data exploration techniques [6] but it is not apparent how they could be applied to the market basket problem. Previous work used PCA for quantitative association rules for estimating errors [7], as well as for load balancing parallel association rule mining mechanisms [8]. Recently, Cadez, Smyth and Mannila [9] examined the use of mixture models for profiling the transaction data of individuals. Here, we will suggest a model of consumer spending behavior originated from (social or cultural) events. These lend themselves to the use of PCA.

II. MODELING CONSUMER SPENDING PATTERNS

A market is based on supply and demand. Shops are typical market places where suppliers provide products to satisfy the demand of consumers. We conjecture that the demand of consumers typically depend on (social or cultural) events. For example, almost every family participates in the "breakfast" event, in which there is a natural distribution of products desired for those events. Certain events are seasonally, like Christmas, while other events may recur like breakfast. Each household will require certain amount of products to be bought in order to satisfy the participation of the events by the household. We assume that each event has a (quasi-stationary) distribution of how likely certain product is bought. For example, suppose there are three products: bread, butter and hammer. The likelihood of a household to buy bread and butter for the breakfast event is much higher than that to buy beer for breakfast. Hence, the spending pattern is dependent on the nature of the event. Obviously, the spending pattern of each household depend on many other factors but for a large number of households, we assume that the probability distribution $d_e = \langle p(i_1),\ldots,p(i_n)\rangle$ of products $\{i_1,\ldots,i_n\}$ for a particular event will be reflected in the aggregate demand of products. Since there are many events that a household is participating, the aggregate demand $D(P)$ of these households is:

$$D(P) = \sum_{h,e} d_{h,e} .$$

This demand may be reflected in the transaction data, representing the eventual spending pattern. Obviously, there are product substitution effects, etc. However, for simplicity, we assume that the consumer only buys the desired product or not. Hence, one consider that $d_{h,e}$ is basically an influence on the consumption, reflected in the transaction data T_h of a consumer h. In general, the influence of such a distribution $d_{h,e}$ is summarized in some function say $G(.)$. However, for simplicity, the influence is simply modeled as additive components, i.e.:

$$T_h = \sum_e p(e) \times d_{h,e}$$

although more sophisticated models using logit transformation for discrete data can be used. Since the events are those that may induce cross-selling effects, it would be important to discover the events and the associated likely products to buy. Hence, it would be interesting to find:

$$d_e = \sum_h p(h) \times d_{h,e} .$$

Assuming that each household is equally likely to consume (i.e. $p(h)$ is a constant),

$$d_e = \sum_h d_{h,e}$$

However, what we can observe is only T_h. Hence, our task is to discover the underlying distribution d_e.

Figure 1 is the schematic diagram of our suggested transaction generation process. Notice that the discovering process may not be able to identify the original number n of events (i.e. $m \neq n$). Note that it is possible to verify which social event that a product is used for by asking the consumer. It is also possible that a single product is used for a number of social events so that the decision to buy a type of product may be based on some aggregate influence for all events that needed the product.

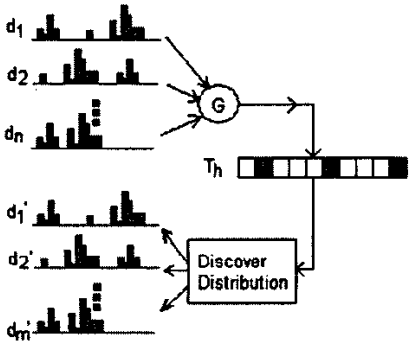

Figure 1: Schematic diagram of discovering the underlying distribution $\{d_1, \ldots, d_n\}$ of consumer spending patterns to participate in different social and/or cultural events.

From a practical point of view, it may not be disastrous even if the discovered distribution differed from the underlying distribution, as long as the distribution found can readily identify products that are sold together in the same transaction for maximizing cross-selling effects. In this vain, it does not matter if two underlying distributions of two different events merged together as long as in the transactions, the consumers bought products to support both events. From the point of view of scientific modeling and from other practical point of views, obtaining the actual underlying distribution is important to predict when certain events are known to occur, for example seasonal events. However, as an initial exploration in this area, we will confine to the case where readily available techniques are used.

III. Principal Component Analysis

Principle component analysis (PCA) is a well-known method in statistical multivariate data analysis and it has been applied widely. First, it was used as a means to reduce the dimensionality of the problem [10] by reducing the number of variables to a few components or latent variables at certain level of accepted information loss. Dimension reduction, however, is not the main objective to apply PCA to the market-basket analysis problem, here.

Second, PCA was used to project data onto a transformed space that is invariant to certain linear transformation, to measure better similarity or dissimilarity of two points. This has been applied in information retrieval and is called latent semantic indexing [11]. Again, this is not our objective to apply PCA to the market-basket analysis problem.

We apply it (i.e. the discovery process in Figure 1 is the PCA) in the market-basket analysis problem (see Figure 2) as a means to discover not the underlying distributions but the significant variables, which are associated positively or negatively to certain events. These events are interpreted as the principal components of the data (i.e. they account for the variability that we observe from the data). The acceptable level of information loss can be considered as noise or underlying uncertainty in the data that cannot be accounted by the principal components or social events. These could be interpreted as a kind of impulsive buying behavior although this type of explanation is anecdote in nature. Figure 2 shows the updated discovery process with a noise component. Note that the noise source is added to the function $G(.)$ since it may not be an additive noise.

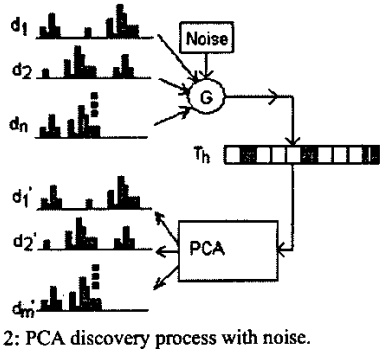

Figure 2: PCA discovery process with noise.

Formally, let us define PCA and interpret the PCA in the context of market basket analysis. PCA assumes that a principal component Z_k is a linear combination of a set of random variables. In this case, these random variables are

categorical variables indicating the presence or absence of an item i_r bought in a transaction T_h. Hence,

$$Z_k = c_{1,k} \times i_1 + \ldots + c_{n,k} \times i_n$$

where $\{c_{i,j}\}$ are the coefficients in the linear combination of the items. In principal component analysis, the number of components is at most n (i.e. $0 < k \leq n$). These principal components can be expressed in matrix notation form, i.e.

$$Z = C \times T_h^t$$

where C is the coefficient matrix $\{c_{j,k}\}$, T^t is the transpose of T. The covariance of Z is:

$$Co\,\mathrm{var}(Z) = C \times \Sigma(T^t) \times C^t$$

where $\Sigma(X)$ returns the square matrix with the main diagonal equals to the eigenvalues $\{\lambda_j\}$ of X and all off-diagonal elements are zero.

In PCA, the coefficients $\{c_{j,k}\}$ are related to the correlation between the observable random variables $\{i_j\}$ and the latent variables $\{Z_k\}$ in the following manner:

$$r(i_j, Z_k) = \frac{c_{j,k}\sqrt{\lambda_j}}{\sqrt{\sigma_{k,k}}}$$

If Z_k are standardized variables, $Covar(Z)$ is the correlation matrix between the latent variables and the correlation between observable random variables $\{i_j\}$ and latent variables $\{Z_k\}$ become:

$$r(i_j, Z_k) = c_{j,k}\sqrt{\lambda_j} \qquad (1)$$

We refer to this correlation as the principal component or PC correlation. The higher the PC correlation $r(i_j, Z_k)$ is the higher the co-occurrence of item i_j with the latent variable Z_k. If there are a number of items or observable random variables correlate with the same principal component, then these items would likely co-correlate or co-occur with each other as well. Hence, we expect that there should be a significant amount of transactions with these co-correlate items. Specifically, if a transaction is classified as being influenced by the principal component or observable event, then the co-correlate items should appear in the transaction.

In PCA, a principal component Z_k is uncorrelated with any other principal component Z_m. From the point of view of modeling, this suggested that social and cultural events are uncorrelated, which is unlikely to be the case. However, for certain general events like those related to a meal and those related to some mechanical work may have little correlation. Hence, this is a strong assumption that PCA made, which are unlikely to be the case in practice.

IV. Initial Exploration

A. Set Up

We carried out a pilot study to examine the potential of using PCA for mining association patterns. A set of 10,000 simulated transaction data is generated by Quest [12] with 20 items. A set of association rules are obtained by the Apriori algorithm [13] with minimum support of 20% and minimum confidence of 30%. In total, 1,157 association rules were discovered and there are 320 unique frequent itemsets. Figure 3 is the scatter diagram of the confidence values and the support values of the mined association rules. The confidence values of all the association rules are larger than or equals to their corresponding support values. Only a number (< 10) of cases (called outliers in the Figure) that their support values and their corresponding confidence values are the same.

For PCA, the standardized variables Z_k are used, which requires the computation of correlations between two categorical variables. Since each variable is a probability of occurrence, the correlation $r(.,.)$ between the presence and absence of item i_r and i_s in a transaction is:

$$r(i_r, i_s) = \frac{p(i_r, i_s) - p(i_r)p(i_s)}{\sqrt{p(i_r)[1 - p(i_r)] \times p(i_s)[1 - p(i_s)]}}$$

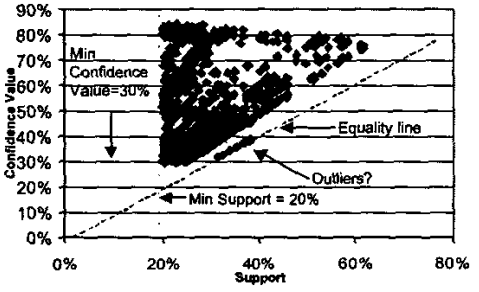

Figure 3: Scatter diagram of the support and confidence values of the set of association rules mined by the Apriori algorithm.

Both the joint probabilities $p(i_r, i_s)$ and individual probabilities $p(i_r)$ are estimated by relative frequency counts. The eigenvalues $\{\lambda_k\}$ of the correlation matrix is shown in Table 1. The information gain IG refers to the amount of data that can be explained by a particular principal component (say i-th), which is:

$$IG_i = \frac{\lambda_i}{\sum_k \lambda_k} \times 100\%.$$

According to the initial results, the largest principal component is 20, which accounts for 12.3% of the data, and which can be expressed as a linear weighted sum of the items as follows:

$$Z_{20} = -0.215X_1 - 0.015X_2 + 0.046X_3 - 0.339X_4 - 0.242X_5 - 0.100X_6 + 0.289X_7 + 0.085X_8 + 0.45X_9 + 0.035X_{10} + 0.190X_{11} + 0.277X_{12} - 0.262X_{13} + 0.319X_{14} - 0.099X_{15} - 0.124X_{16} + 0.357X_{17} + 0.165X_{18} - 0.097X_{19} - 0.037X_{20}$$

The coefficients in the above equation can be converted into a set of correlation values between Z_{20} and individual items, using equation (1), as in Table 2. If we consider those correlation coefficients larger than 20% as strong, then item 7, 9, 11, 12, 14, 17 and 18 are strongly correlated with the largest principal component. This suggest that if a transaction is principally influenced by the largest principal component, then it is likely to observe these correlated items although they may not necessarily simultaneously occur in the same transaction.

Component	Eigenvalue	Information Gain
1	0.876	4.38%
2	0.944	4.72%
3	0.942	4.71%
4	0.881	4.40%
5	0.89	4.45%
6	0.901	4.50%
7	0.955	4.77%
8	0.917	4.58%
9	0.963	4.81%
10	0.861	4.30%
11	0.856	4.28%
12	0.982	4.91%
13	0.837	4.18%
14	0.988	4.94%
15	0.817	4.08%
16	1.007	5.03%
17	0.809	4.04%
18	1.034	5.17%
19	1.076	5.38%
20	**2.465**	**12.32%**

Table 1: Eigenvalues of the principal components.

I_1	I_2	I_3	I_4	I_5
-30%	0%	10%	-50%	-40%
I_6	I_7	I_8	I_9	I_{10}
-20%	**50%**	10%	**70%**	10%
I_{11}	I_{12}	I_{13}	I_{14}	I_{15}
30%	**40%**	-40%	**50%**	-20%
I_{16}	I_{17}	I_{18}	I_{19}	I_{20}
-20%	**60%**	**30%**	-20%	-10%

Table 2: Correlation between items and the largest principal component.

B. Comparison

To compare whether there are any relationships between mined association rules and the principal components, we examine whether the strong PC correlations for a given component match with any of the association rules and related frequent itemsets, since we conjecture that the strong PC correlations suggest cross buying behavior. Related to the data generation model, our assumption is that the buyer driven by some social or cultural event needs to buy (multiple) items to participate in the event. Hence, the transaction would reflect the co-occurrences of items supporting specific events. These co-occurrences would be related to the frequent itemsets during the association rule mining process and individual association rules are simply the ratio of the probability of two itemsets where one itemset is a subset of the other, i.e. confidence value *conf(.)* of the association rule for itemset $X \rightarrow Y$ is:

$$conf(X \rightarrow Y) = \frac{p(X \cup Y)}{p(X)}$$

where X and Y are disjoint (for the Apriori algorithm). Therefore, an initial comparison between the principal components and the mining of association rules is to examine whether the itemsets of both algorithms are the same. Since the frequent itemsets of the association rules are always correct, they are used as a reference for comparison.

For PCA, there is a need to define what are strong PC-correlations since those are considered to be co-occurring for a given event. Instead of defining a threshold for strong PC correlations, we examine how the matching performance varies with different level of threshold values so that the tendency and characteristics of matching in relation to the different threshold values can be examined.

For matching, we expect that as long as all the items in a frequent itemset has strong correlations, then there is a match (i.e. for all items i in th itemset X, $r(i, Z)$ is strong for some principal component Z), since it is possible to recover the frequent itemset by further analysis. Since there is no threshold to define strong correlation, for the itemset X to be discovered by PCA, the threshold has to be below or equals to the minimum PC correlation value of all the items in X for the principal component Z. Since there are more than one principal component, if any principal component, can match with the itemset X, then X can be discovered by further analysis. Hence, the threshold can set to the maximum of all the different minimum PC correlation values of the itemsets of different principal components. We call this threshold below which the itemset X can be discovered by PCA the min-max correlation value for itemset X and formally, it is defined as:

$$\min - \max(X) \equiv \max_{Z} \left\{ \min_{i \in X} \{ r(i, Z) \} \right\} \quad (2)$$

where Z is a principal component.

Figure 4 shows the min-max correlation values against the maximum confidence value of the frequent itemsets. This maximum confidence value is the maximum confidence values of all association rules, $X \rightarrow Y$, such that itemset $W = X \cup Y$, i.e.

$$\max - conf(W) \equiv \max_{W} \{conf(X \rightarrow Y) \mid W = X \cup Y\} \quad (3)$$

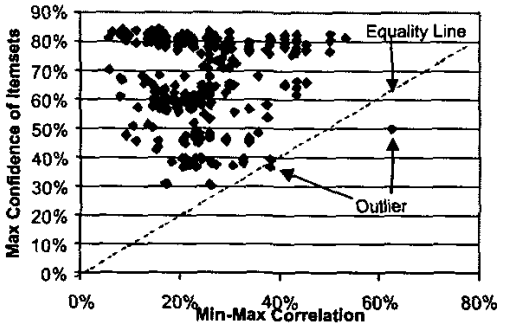

Figure 4: Scatter diagram of the min-max correlation values (definition 2) of itemsets against their corresponding maximum confidence value (definition 3) derived from related association rules (see text).

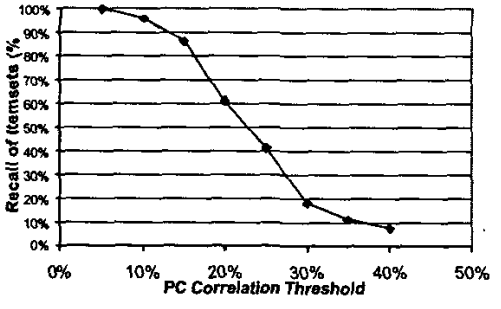

Figure 5: The effect of PC correlation threshold on the recall of itemsets.

The maximum was used because if the itemset W leads to any quality association rules, W should be retained rather than filtered. According to Figure 4, in almost all cases, except 2, that *max-conf(W)* > *min-max(W)*. Hence, we expect the discovered itemsets to have a better confidence value than the min-max correlation value, which may be used as a kind of approximate lower bound to discover quality association rules.

Figure 5 shows the impact of filtering PC correlation values using different thresholds on the recall of itemsets as discovered by the association rules. As the threshold increases, the recall dropped dramatically. Whether this impact is desirable depends on whether the filtered itemsets are low quality, i.e. have low *max-conf* values.

We observe the effects of setting different PC correlation threshold in Figure 6 on filtering the (quality) itemsets. As the threshold value increases the minimum *max-conf* values for all the itemsets increases semi-monotonically and drastically after 35%, and the maximum *max-conf* values for all the itemsets dropped by 1% (from 84% to 83%). This shows that certain quality association rules are retained with increasing threshold value. In addition, the average *max-conf* values steadily converge towards the maximum *max-conf* values as the threshold increases.

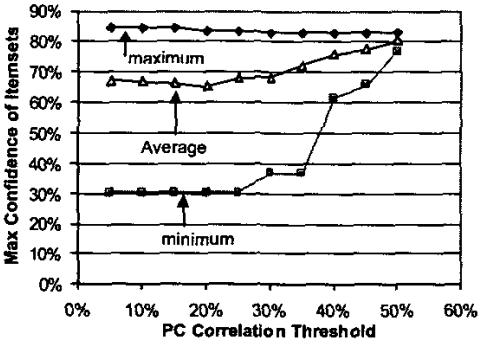

Figure 6: PC correlation threshold against the itemsets discovered.

Figure 7 shows that the amount of PC correlation values that are filtered against the quality of the itemsets measured by the *max-conf* values set by different threshold. With just 10% of the top correlation values retained, the itemsets leading to the quality association rules can be found using PCA. In this case, since there are 400 (i.e. 20 × 20) PC correlations, 10% represents retaining only 40 PC correlation coefficients and only 2 correlation coefficients per principal component.

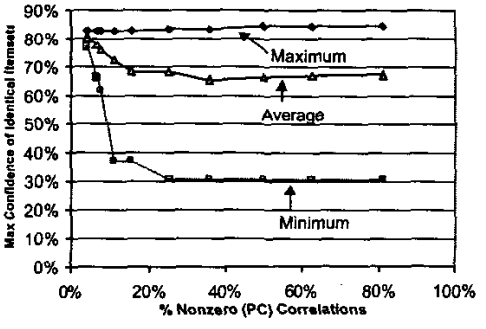

Figure 7: The amount of nonzero PC correlations filtered against the quality of the itemsets found, measured by the *max-conf* values.

Figure 8 shows the size of itemsets discovered by PCA against the different percentages of nonzero PC correlations of all the principal components. The general trend is that the less the amount of percentages the smaller the itemset size. In particular, even though the percentage of nonzero PC correlations is close to 0%, the maximum item size discovered by PCA can still be as large as 4. Since certain principal components have no nonzero PC correlations when the

percentage is close to 0%, these principal components can be discarded without any further data exploration.

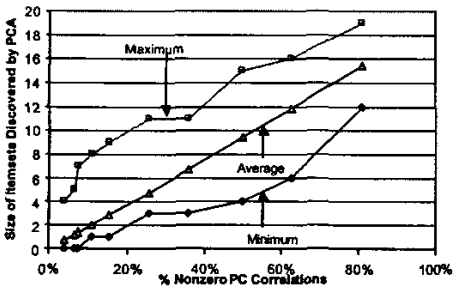

Figure 8: The maximum, minimum and average number of nonzero PC correlations per principal component against the different percentages of nonzero PC correlations for all the principal components.

Another encouraging aspect of using PCA is that the average number of itemset size is only 15 compared with 20 because some of the PC correlations are zeros inherently. When the PC correlation threshold is set at 0.0, the recall of itemsets is 100%. Hence, there may be some potential in saving processing speed using less number of items.

V. Summary

In this paper, we have explored the use of finite mixture densities to model the consumer spending patterns. Products bought for specific social and/or cultural events were considered to be the underlying driving forces of cross selling effects as multiple related items are needed for consumers to participate into those events. We applied the principal component analysis (PCA) to discover these events as principal components, taking a simplistic view of the discovery process, for this initial exploration. We used a set of 10,000 transaction data generated by Quest [12] to examine how PCA discovered co-occurring items may relate to association rules mined using the Apriori algorithm.

Our initial understanding is that the frequent itemsets of association rules may relate to the co-occurring items discovered by PCA. These co-occurring items are thought to be the strong correlations between the principal component and the specific item. Instead of using a threshold to define strong correlations, we use a novel measure called the min-max correlation to illustrate graphically how PCA mining of patterns relate to the frequent itemsets discovered by the Apriori algorithm. Our initial study show that as the correlation threshold is increased, the average, minimum and maximum confidence values of the itemsets discovered by the Apriori algorithm and the PCA converges to high confidence value (around 80%). However, this process would loose some high confidence value rules. Nevertheless, certain high confidence and nontrivial association rules related to the itemsets remain. Further work is necessary to demonstrate the extend of the utility of PCA.

Acknowledgement

This research is supported in part by project funding for the postgraduate study.

Reference

[1] Agrawal, R., T. Imielinski and A Swami. Mining association rules between sets of items in large database. *In Proceedings of the 1993 International Conference on Management of Data (SIGMOD 93),* pages 207-216, May 1993.

[2] Agrawal, R. and R. Srikant. *Fast algorithms for mining association rules in large databases. In VLDB' 94,* September 1994.

[3] Russell, G.J. and A. Petersen. Analysis of cross category dependence in market basket selection, *Journal of Retailing.* 67(3): 367-392, 2000.

[4] Meo, R. Theory of dependence values, *ACM Trans. on Database Systems*, 25(3): 380-406, 2000.

[5] Bayardo Jr., R. Efficiently mining long patterns from databases, *In Proc. 1998 International Conference on Management of Data (SIGMOD 98),* pages 85-93, 1998.

[6] Bryan F.J. Manly. *Multivariate Statistical Methods* A *Primer*, Chapman and Hall, 1986

[7] F. Korn, A. Labrinidis, Y. Kotidis, C. Faloutsos, A. Kaplunovich, and D. Perkovic. Quantifiable data mining using principal component analysis, *CS-TR-3754 and UMIACS-TR-97-13 technical reports,* February 1997.

[8] Manning, A.M. and J.A. Keane. Inducing load balancing and efficient data distribution prior to association Rule Discovery in a Parallel Environment, *In Proc. European Conference on Parallel Processing,* pages 1460-1463, 1999.

[9] Cadez, I.V., Smyth P. and Mannila H. Probabilistic modeling of transaction data with applications of profiling, visualization and prediction. *In Proc. ACM KDD Conference*, pages 37-46, 2001,

[10] Levin, A.U., Leen T.K. and Moody, J.E. Fast pruning using principal components, *Advances in Neural Information Processing Systems*, 6: 35-42, 1994.

[11] Aggarwal, C.C. On the effects of dimensionality reduction on high dimensional similarity search, *In ACM PODS Conference*, 2001.

[12] Agrawal, R. M. Mehta, J. Shafer, R. Srikant, A. Arning and T. Bollinger. The Quest data mining system, *In Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining*, pages 244-249, 1996.

[13] Han, J. M Kamber. Data mining: concepts and techniques. *Morgan Kaufmann Publishers*, 2001.