Demiriz, A., Ertek, G., Atan, T., and Kula, U. (2010) "Re-mining Positive and Negative AssociationMining Results" P. Perner (Ed.): Advances in Data Mining. Applications and Theoretical Aspects, 10th Industrial Conference, ICDM 2010, Berlin, Germany, July 12-14, 2010. Proceedings. LNAI 6171, pp. 101–114.

Re-mining Positive and Negative Association Mining Results

Ayhan Demiriz[1], Gurdal Ertek[2], Tankut Atan[3], and Ufuk Kula[1]

[1]Sakarya University

Sakarya, Turkey

[2]Sabanci University

Istanbul, Turkey

[3]Isik University

Istanbul, Turkey

Re-mining Positive and Negative Association Mining Results

Ayhan Demiriz[1], Gurdal Ertek[2], Tankut Atan[3], and Ufuk Kula[1]

[1] Sakarya University
Sakarya, Turkey
{ademiriz,ufukkula}@gmail.com
[2] Sabanci University
Istanbul, Turkey
ertekg@sabanciuniv.edu
[3] Isik University
Istanbul, Turkey
tatan@isikun.edu.tr

Abstract. Positive and negative association mining are well-known and extensively studied data mining techniques to analyze market basket data. Efficient algorithms exist to find both types of association, separately or simultaneously. Association mining is performed by operating on the transaction data. Despite being an integral part of the transaction data, the pricing and time information has not been incorporated into market basket analysis so far, and additional attributes have been handled using quantitative association mining. In this paper, a new approach is proposed to incorporate price, time and domain related attributes into data mining by re-mining the association mining results. The underlying factors behind positive and negative relationships, as indicated by the association rules, are characterized and described through the second data mining stage *re-mining*. The applicability of the methodology is demonstrated by analyzing data coming from apparel retailing industry, where price markdown is an essential tool for promoting sales and generating increased revenue.

1 Introduction

Association mining is a data mining technique in which the goal is to find rules in the form of $X \Rightarrow Y$, where X and Y are two non-overlapping sets of items or events, depending on the domain. A rule is considered as significant if it is satisfied by at least a percentage of cases specified beforehand (`minimum support`) and its confidence is above a certain threshold (`minimum confidence`). Conventional association mining considers "positive" relations as in the rule $X \Rightarrow Y$. However negative associations such as $X \Rightarrow \neg Y$, where $\neg Y$ represents the negation (absence) of Y, might also be discovered through association mining.

Association mining has contributed to many developments in a multitude of data mining problems. Recent developments have positioned the association mining as one of the most popular tools in retail analytics, as well [1]. Traditionally,

P. Perner (Ed.): ICDM 2010, LNAI 6171, pp. 101–114, 2010.

association mining generates positive association rules that reveal complementary effects. In other words, the rules suggest that purchasing an item can generate sales of other items. Association mining can also be used to find so-called "Halo effects", where reducing the price of an item can entice and increase the sales of another item. Although positive associations are an integral part of retail analytics, negative associations are not. However negative associations are highly useful to find out the substitution effects in a retail environment. Substitution means that a product is purchased instead of another one.

There have been numerous algorithms introduced to find positive and negative associations since the pioneering work of Agrawal et al. [2]. Market basket analysis is considered as a motivation and a test bed for these algorithms. Since the price data are readily available in the market basket data, one might expect to observe the usage of price data in various applications. Conceptually quantitative association mining [3,4] can handle pricing data and other attribute data. However pricing data have not been utilized before as a quantitative attribute except in [5], which explores a solution with the help of singular value decomposition. Quantitative association mining is not the only answer to analyze the attribute data by conventional association mining. Multidimensional association mining [4] is also a methodology that can be adapted in analyzing such data. Inevitably, the complexity of association mining will increase with the usage of additional attribute data where there might be both categorical and quantitative attributes in addition to the transaction data [6]. Even worse, the attribute data might be less sparse compared to transaction data.

The main objective of this paper is to develop an efficient methodology that enables incorporation of attribute data (e.g. price, category, sales timeline) to explain both positive and negative item associations. Positive and negative item associations indicate the complementarity and substitution effects respectively. To the best of our knowledge, there exists no methodological research in data mining literature that enables such a multi-faceted analysis to be executed efficiently and is proven on real world attribute data. A practical and effective methodology is developed to discover how price, item, domain and time related attributes affect both positive and negative associations by introducing a new data mining process.

As a novel and broadly applicable concept, we define *data re-mining* as mining a newly formed data from the results of an original data mining process. Aforementioned newly formed data will contain additional attributes on top of the original data mining results. These attributes in our case will be related to price, item, domain and time. Our methodology combines pricing as well as other information with the original association mining results within the framework of a new mining process. We thereby generate new rules to characterize, describe and explain the underlying factors behind positive and negative associations. Remining is a different process from post-mining where the latter only summarizes the data mining results. For example visualizing the association mining results [7] could be regarded as a post-mining activity. Our methodology extends and generalizes post-mining process.

Our work contributes to the field of data mining in three ways:

1. We introduce a new data mining concept and its associated process, named as *Re-Mining*, which enables an elaborate analysis of both positive and negative associations for discovering the factors and explaining the reasons for such associations.
2. We enable the efficient inclusion of price data into the mining process in addition to other attributes of the items and the application domain.
3. We illustrate that the proposed methodology is applicable to real world data from apparel retailing.

The remainder of the paper is organized as follows. In Section 2, an overview of the basic concepts in related studies is presented through a concise literature review. In Section 3, Re-Mining is motivated, defined, and framed. The methodology is put into use with apparel retail data and its applicability is demonstrated in Section 4. In Section 5, the limitations of the quantitative association regarding the retail data used in this paper are shown. Finally, Section 6 summarizes our work and discusses the future directions.

2 Related Literature

Data mining can simply be defined as extracting knowledge from large amounts of data [4]. An extended definition additionally requires that findings are meaningful, previously unknown and actionable [8]. Interpreting the results is an essential part of the data mining process and can be achieved through the post-mining analysis of multi-dimensional association mining results.

Quantitative and multi-dimensional association mining techniques can integrate attribute data into the association mining process where the associations among these attributes are also found. However, these techniques introduce significant additional complexity, since association mining is carried out with the complete set of attributes rather than just the market basket data. In the case of quantitative association mining quantitative attributes are transformed into categorical attributes through discretization, transforming the problem into multi-dimensional association mining with only categorical attributes. This is an *NP-Complete* problem as shown in [6], with the exponentially increasing running time as the number of additional attributes increases linearly.

In re-mining *single dimensional* rules are generated and then expanded with additional attributes. Multi-dimensional association mining, on the other hand, works directly towards the generation of multi-dimensional rules. It relates all the possible categorical values of all the attributes to each other. In our methodology, the attribute values are investigated only for the positively and negatively associated item pairs, with much less computational complexity.

2.1 Negative Association Mining

The term association mining is generally used to represent positive association mining. Since positive association mining has been studied extensively,

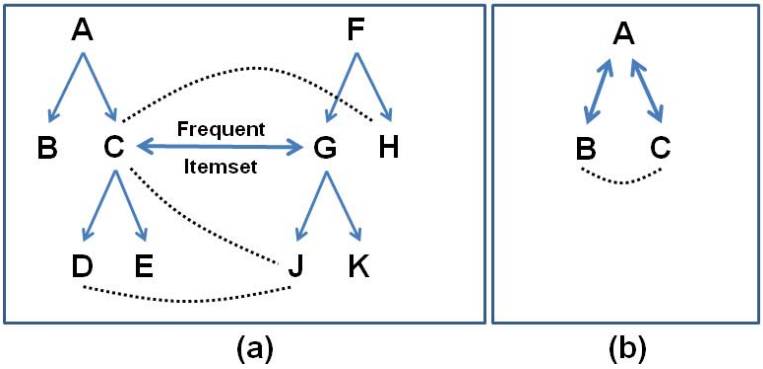

Fig. 1. (a) Taxonomy of Items and Associations [9]; (b) Indirect Association

we limit ourselves to review some of the approaches described in the literature for finding negative associations. One innovative approach utilizes the domain knowledge of item hierarchy (taxonomy), and seeks negative association between items in a pairwise way [9]. Authors in [9] propose the interestingness measure (RI) based on the difference between expected support and actual support: $RI = \frac{E[s(XY)]-s(XY)}{s(X)}$. A minimum threshold is specified for the interestingness measure RI besides the minimum support threshold for the candidate negative itemsets. Depending on the taxonomy (e.g. Figure 1(a)) and the frequent itemsets, candidate negative itemsets can be generated. For example, assuming that the itemset $\{CG\}$ is frequent in Figure 1(a), the dashed curves represent some candidate negative itemsets.

In [10], negative associations are found through indirect associations. Figure 1(b) depicts such an indirect association $\{BC\}$ via item A. In Figure 1(b) itemsets $\{AB\}$ and $\{AC\}$ are both assumed to be frequent, whereas the itemset $\{BC\}$ is not. The itemset $\{BC\}$ is said to have an *indirect association* via the item A and thus is considered as a candidate negative association. Item A in this case is called as a *mediator* for the itemset $\{BC\}$. Just like the aforementioned method in [9], indirect association mining also uses an interestingness measure -*dependency* in this case- as a threshold. Indirect mining selects as candidates the frequent itemsets that have strong dependency with their mediator.

Even though both methods discussed above are suitable for retail analytics, the approach in [10] is selected in our study to compute negative associations due to convenience of implementation.

2.2 Quantitative and Multi-dimensional Association Mining

The traditional way of incorporating quantitative data into association mining is to discretize (categorize) the continuous attributes. An early work by Srikant and Agrawal [3] proposes such an approach where the continuous attributes are

first partitioned and then treated just like categorical data. For this, consecutive integer values are assigned to each adjacent partition. In case the quantitative attribute has few distinct values, consecutive integer values can be assigned to these few values to conserve the ordering of the data. When there is not enough support for a partition, the adjacent partitions are merged and the mining process is rerun. Although [3] emphasizes rules with quantitative attributes on the left hand side (antecedent) of the rules, since each partition is treated as if it were categorical, it is also possible to obtain rules with quantitative attributes on the right hand side (consequent) of the rules.

A more statistical approach is followed in [11] for finding association rules with quantitative attributes. The rules found in [11] can contain statistics (mean, variance and median) of the quantitative attributes.

As discussed earlier, re-mining does not investigate every combination of attribute values, and is much faster than quantitative association mining. In Section 5, quantitative association mining is also carried out for the sake of completeness.

Finally, in [5], where the significant ratio rules are found to summarize the expenses made on the items. An example of ratio rule would be "*Customers who buy bread:milk:butter spend 1:2:5 dollars on these items*" [5]. This is a potentially useful way of utilizing the price data for unveiling the hidden relationships among the items in sales transactions. According to this approach, one can basically form a price matrix from sales transactions and analyze it via singular value decomposition (SVD) to find positive and negative associations. Ensuring the scalability of SVD in finding the ratio rules is a significant research challenge.

2.3 Learning Association Rules

In [12,13] a framework is constructed based on the idea of learning a classifier to explain the mined results. However, the described framework considers and interprets only the positive association rules, whether they are interesting or not. The approach obligates human intervention for labeling the generated rules as interesting or not. The framework in [12,13] is the closest work in the literature to our re-mining approach. However our approach is very different in the sense that it includes negative associations and is suitable for the automated rule discovery to explain the originally mined results. Unlike in [12,13], our approach is applied to a real world dataset.

Based on correlation analysis, authors in [14] propose an algorithm to classify associations as positive and negative. The learning is only used on correlation data and the scope is narrowly determined to label the associations as positive or negative.

3 The Methodology

In this section we introduce the proposed methodology, which transforms the post-mining step into a new data mining process. The re-mining algorithm consists of following steps.

1. *Perform association mining.*
2. *Sort the items in the 2-itemsets.*
3. *Label the item associations accordingly and append them as new records.*
4. *Expand the records with additional attributes for re-mining.*
5. *Perform exploratory, descriptive, and predictive re-mining.*

Fig. 2. Re-Mining Algorithm

Potentially, re-mining algorithm can be conducted in explanatory, descriptive, and predictive manners. Re-mining can be considered as an additional data mining step of KDD process [8]. We define *re-mining* process as combining the results of an original data mining process with a new set of data and then mining the newly formed data again. Conceptually, re-mining process can be extended with many more repeating steps since each time a new set of the attributes can be introduced and a new data mining technique can be utilized. However the re-mining process is limited to only one additional data mining step in this paper. In theory, the new mining step may involve any appropriate set of data mining techniques.

Data mining does not require any pre-assumptions (hypotheses) about the data. Therefore the results of data mining may potentially be full of surprises. Making sense of such large body of results and the pressure to find surprising insights may require incorporating new attributes and subsequently executing a new data mining step, as implemented in re-mining. The goal of this re-mining process is to explain and interpret the results of the original data mining process in a different context, by generating new rules from the consolidated data. The results of re-mining need not necessarily yield an outcome in parallel with the original outcome. In other words, if the original data mining yields for example frequent itemsets, it is not expected from re-mining to output frequent itemsets again.

The main contribution of this paper is to introduce the re-mining process to discover new insights regarding the positive and negative associations. However the usage of re-mining process is not limited to this. We can potentially employ the re-mining process to bring further insights to the results of any data mining task.

The rationale behind using the re-mining process is to exploit the domain specific knowledge in a new step. One can understandably argue that such background knowledge can be integrated into the original data mining process by introducing new attributes. However, there might be certain cases that adding such information would increase the complexity of the underlying model [6], and diminish the strength of the algorithm. To be more specific, it might be necessary to find attribute associations when the item associations are also present, which requires constraint-based mining [15]. Re-mining may help with grasping the causality effects that exist in the data as well, since the input of the causality models may be an outcome of the another data mining process.

4 Case Study

In this section the applicability of the re-mining process is demonstrated through a real world case study that involves the store level retail sales data originating from an apparel retail chain. In our study, we had access to complete sales, stock and transshipment data belonging to a single merchandise group (men's clothes line) coming from the all stores of the retail chain (over 200 of them across the country) for the 2007 summer season. Throughout the various stages of the study, MS SQL Server, SAS, SPSS Clementine, Orange, and MATLAB software packages have been used as needed.

First, the retail data and its data model used in this study are described. Then the application of the re-mining methodology on the given dataset is presented. Re-mining reveals the profound effect of pricing on item associations and frequent itemsets, and provides insights that the retail company can act upon.

4.1 Retail Data

A typical product hierarchy for an apparel retailer displays the following sequence: *merchandise group*, *sub-merchandise group*, *category*, *model*, and *SKU*. The products at the Stock Keeping Unit (*SKU*) level are sold directly to the customers. At the *SKU* level, each color and size combination that belongs to a *model* is assigned a unique SKU number. A *category* is composed of similar models, e.g. long sleeve shirts. A *merchandise group* can represent the whole product line for a gender and age group, e.g. men's clothes line. A *sub-merchandise group* divides this large group. Notice that this particular product hierarchy is just a typical representation and the hierarchy may vary from one firm to another.

Since the SKU level store data exhibit high variability, the dataset is aggregated at the model level, the immediate parent of the SKU level. Sales transaction data consist of a collection of rows generated by a sale, that includes an item numbers, and their prices, a transaction identifier, and a time stamp. Positively and negatively associated item pairs can thus be found using transactional data. For apparel products, price is an important factor influencing the purchasing decisions of consumers, and markdown management (planning the schedule and price levels of discounts) is an essential activity for increasing the revenue of an apparel chain. Consequently, pricing is an important driver of the multiple-item sales transactions and is highly relevant to association mining activities in apparel retailing.

Out of the 710 models available in the dataset, the top 600 models have been selected according to sales quantities. Most of the sales consist of single-item purchases. There exist 2,753,260 transactions and 4,376,886 items sold. Technically it is hard to find positive associations in sparse data and the sparsity of the data is very high with $\sim$99.74%. Although single-item transactions could be removed from a traditional association mining task, they are included in the case study to observe the effect of the pricing. In other words, inclusion of single item purchases does not change the association mining statistics (support values), yet enables accurate calculation of the values of additional attributes.

For example, for calculating the average price of an item across the season, one will obtain a more accurate statistic when single item transactions are also included.

4.2 Conventional Association Mining

As the first step of re-mining methodology, conventional association mining has been conducted. Apriori algorithm was run with a minimum support count of 100 to generate the frequent itemsets. All the 600 items were found to be frequent. In re-mining, only the frequent 2-itemsets have been investigated and 3930 such pairs have been found. Thus frequent itemsets were used in the analysis, rather than association rules. The top-5 frequent pairs given in Table 1 have support counts of 22131, 17247, 17155, 14224, and 11968 respectively, within the 2,753,260 transactions. We utilize the retail data in transactional format as known in conventional association mining. Item names are replaced by the alphabet letters for brevity.

Table 1. Top-5 Frequent Pairs

Item 1	Item 2	S Count
A	B	22131
B	F	17247
A	E	17155
B	E	14224
C	B	11968

The frequent pairs were then used in finding the negatively related pairs via indirect association mining. Negative relation is an indicator of product substitution. Implementing indirect association mining resulted in 5,386 negative itemsets, including the mediator items. These itemsets were reduced to a set of 2,433 unique item pairs when the mediators were removed. This indeed shows that a considerable portion of the item pairs in the dataset are negatively related via more than one mediator item.

4.3 Re-mining the Expanded Data

Following conventional association mining, a new data set E^* was formed from the item pairs and their additional attributes A^* for performing exploratory, descriptive and predictive re-mining. In this paper, we only illustrate descriptive re-mining by using decision tree analysis and a brief exploratory re-mining example due to space considerations. As a supervised classification method, decision tree approach usually requires the input data in a table format, in which one of the attributes is the class label. The type of association, positive ('+') or negative ('-'), was selected as the class label in our analysis.

An item pair can generate two distinct rows for the learning set - e.g. pairs AB and BA, but this representation ultimately yields degenerate rules out of

learning process. One way of representing the pair data is to order (rank) items in the pair according to a sort criterion. In the case study, sort attribute was selected as the price, which marks the items as higher and lower priced items, respectively.

For computing price-related statistics (averages and standard deviations) a price-matrix was formed out of the transaction data. The price-matrix resembles the full-matrix format of the transaction data with the price of the item replacing the value of **1** in the full-matrix. The price-matrix was normalized by dividing each column by its maximum value, enabling comparable statistics. A price value of 0 in the price-matrix means that the item is not sold in that transaction. The full price-matrix has the dimensions $2,753,260 \times 600$.

Besides price related statistics such as minimum, maximum, average and standard deviations of item prices (MinPriceH, MaxPriceH, MinPriceL, MaxPriceL, AvgPriceH_H1_L0, ..., StdDevPriceH_H1_L0, ...), attributes related with time and product hierarchy were appended, totaling to 38 additional attributes. All the additional attributes have been computed for all sorted item-pairs through executing relational queries.

Once the new dataset is available for the re-mining step, any decision tree algorithm can be run to generate the descriptive rules. Decision tree methods such as C5.0, CHAID, CART are readily available within data mining software in interactive and automated modes. An interactive decision tree is an essential tool for descriptive discovery, since it grows with user interaction and enables the verification of the user hypotheses on the fly.

If decision tree analysis is conducted with all the attributes, support count related attributes would appear as the most significant ones, and data would be perfectly classified. Therefore it is necessary to exclude the support item attributes from the analysis to arrive a conclusive description of the data.

The C5.0 algorithm was executed in automated mode with the default settings, discovering 53 rules for the '-' class, and 11 rules for the '+' class (the default class). One example of the rule describing the '-' class is as follows: **If** StartWeekH $>$ 11 **and** AvgPriceL_H0_L1 $>$ 0.844 **and** CategoryL $=$ 0208 **and** CorrNormPrice_HL $\leq$ 0.016 **Then** '-'. This rule reveals that when

- the higher priced is sold after 11^{th} calendar week *and*
- average (normalized) price of lower priced item is greater than 0.844 (that corresponds to roughly maximum 15% discount on average) when it is sold and the higher priced item is not sold *and*
- category of lower priced item is "0208" *and*
- correlation coefficient between normalize prices of higher and lower price items is less than or equal to 0.016 then the target class is '-'.

An example of '+' class rule is that **If** MaxPriceH $\leq$ 14.43 **and** StdDevPriceH_H1_L0 $\leq$ 0.05 **Then** '+'. Another example of the '+' class rule is the rule **If** LifeTimeH $>$ 23 **and** LifeTimeL $<=$ 21 **and** MaxPriceH $>$ 12.21 **and** CorrNormPrice_HL $>$ 0.003 **Then** '+'. This rule basically emphasizes on the lifetime of the higher priced item and its maximum price with respect to item association.

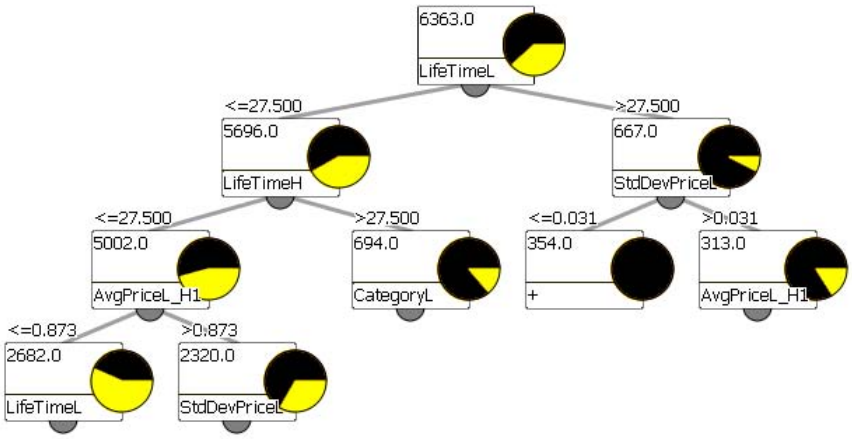

Fig. 3. An Illustrative Decision Tree Model in Re-Mining

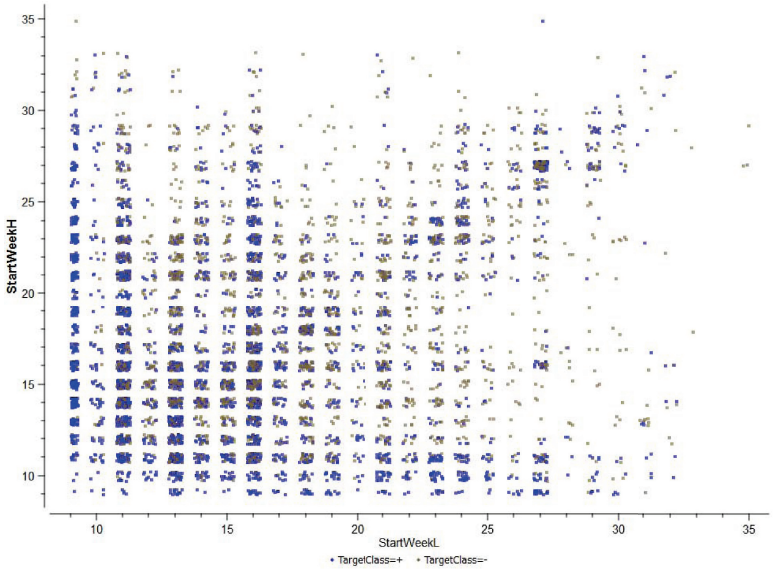

Fig. 4. Exploratory Re-Mining Example: Analyzing Item Introduction in Season

The decision tree in Figure 3 is obtained by interactively pruning the full-tree. It can be observed that LifeTimeL, the lifetime of the lowered price item (in weeks) is the most significant attribute to classify the dataset. If LifeTimeL greater than or equal to 27.5 weeks,the item pair has much higher chance of having positive association.

As seen from Figure 4, positive ("Target Class=+") and negative ("Target-Class=-") associations can be separated well at some regions of the plot. For

example, if StartWeekL is around 5 (calendar week) then the associations are mainly positive regardless of the starting week of the higher priced item. In addition, a body of positive associations up to 'StartWeekL=15' can be seen in Figure 4. It means that there is a high possibility of having a positive association between two items when the lower priced item is introduced early in the season. We can maybe conclude that since basic items are usually introduced early in the season, there is a big chance of having a positive association between two items when the lower priced item is a basic one but not a fashion one.

Similar conclusions can be attained when the maximum prices of items considered are compared in Figure 5. It can easily be seen that negative associations usually occur when the maximum prices of items are higher than 25. It should be noted that the basic items may be priced below 20-25 range. Thus it is very likely that many of the fashion items might have negative associations between themselves.

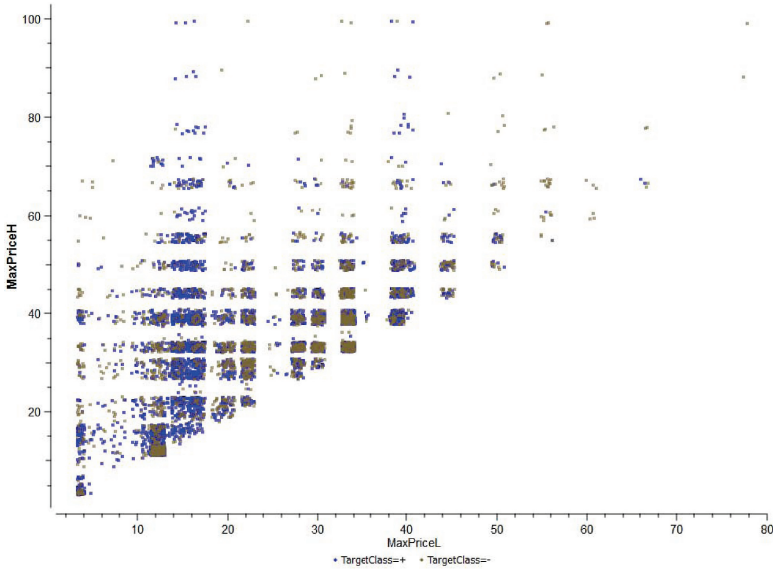

Fig. 5. Exploratory Re-Mining Example: Effect of the Maximum Item Price

5 Comparison with Quantitative Association Mining

Additional attribute data used can also be incorporated through quantitative association mining, as an alternative to re-mining. This type of analysis has also been conducted to illustrate its limitations on analyzing retail data. Quantitative association mining has been studied extensively in the literature and some of the major applications like [3] were reviewed in Section 2.2. After an initial analysis of the price data, it was observed that there are not too many price levels for the

products. Therefore a straight-forward discretization, which does not require a complex transformation, exists.

Notice that one of the most important steps in the quantitative association mining is the discretization step. Instead of utilizing a more complex quantitative association mining, we take the liberty to conjoin the item and the price information into a new entity to conduct our quantitative association analysis in a simplified way.

Two main seasons in apparel retailing, winter and summer have approximately equal length. As a common business rule, prices are not marked down too often. Usually, at least two weeks pass by between subsequent markdowns. Thus, there exist a countable set of price levels within each season. There might be temporary sales during the season, but marked down prices remain the same until the next price markdown. Prices are set at the highest level in the beginning of each season, falling down by each markdown. If the price data of a product is normalized by dividing by the highest price, normalized prices will be less than or equal to 1. When two significant digits are used after the decimal point, prices can be easily discretized. For example, after an initial markdown of 10%, the normalized price will be 0.90. Markdowns are usually computed on the original highest price. In other words, if the second markdown is 30% then the normalized price is computed as 0.70.

Table 2. Top-10 Frequent Pairs for the Quantitative Data

Pair ID	Item 1	Item 2	Sup. Count
1	A_0.90	B_0.63	14312
2	A_0.90	E_0.90	10732
3	B_0.63	E_0.90	8861
4	C_0.90	B_0.63	7821
5	A_1.00	B_0.70	7816
6	A_1.00	E_1.00	6377
7	B_0.63	F_0.90	5997
8	B_0.70	E_1.00	5344
9	B_0.63	F_0.78	5318
10	D_0.90	B_0.63	4402

After using this discretization scheme, 3,851 unique product-normalized price pairs have been obtained for the 600 unique products of the original transaction data. Each product has 6 price levels on the average. The highest number of price levels for a product is 14 and the lowest number of price levels is 2. This shows that markdowns were applied to all the products and no product has been sold at its original price throughout the whole season. Technically, a discretized price can be appended to a corresponding product name to create a new unique entity for the quantitative association mining. For example appending 0.90 to the product name 'A' after an underscore will create the new entity 'A_0.90'. One

can easily utilize the conventional association mining to conduct a discretized quantitative association mining after this data transformation (discretization).

The top-10 frequent pairs and their support counts are depicted in Table 2 where item IDs are masked. As can be observed from Table 2, a large portion of the frequent purchases occurs at the discounted prices. Among the top-10 frequent item pairs, only the sixth one has full prices for both of the items. The remaining pairs are purchased at marked down (discounted) prices.

The retail company does not allow price differentiations by locations at any give time. In other words, an item will have the same price across all the stores at any given time. Quantitative association mining can identify negative associations between items for value combinations that actually never occur. For example, even though an item A is sold only at the normalized price of 0.70 in the time interval that B is sold, quantitative association mining can still suggest other price combinations of A and B, such as (A_1.00, B_1.00) as negative item pairs. Thus, many item-price combinations will have negative associations due to the nature of the business and the way quantitative association mining operates, yielding misleading results.

Even though both positive and negative quantitative association mining can be run on any given dataset conceptually, it is not guaranteed to yield useful outcomes. Alternatively, re-mining operates only on the confirmed positive and negative quantitative associations and does not exhibit the discussed problem.

6 Conclusion

A framework, namely re-mining, has been proposed to enrich the original data mining process with a new set of data and an additional data mining step. The goal is to describe and explore the factors behind positive and negative association mining, and to predict the type of associations based on attribute data. It is shown that not only categorical attributes (e.g. category of the product) but also quantitative attributes such as price, lifetime of the products in weeks and some derived statistics, can be included in the study while avoiding NP-completeness.

The framework has been demonstrated through a case study in apparel retail industry. Only descriptive and a brief exploratory re-mining have been performed in this paper. Predictive re-mining can also be conducted, and this is planned as a future study. Our case study has revealed some interesting outcomes such as the negative associations are usually seen between fashion items and the price of an item is an important factor for the item associations in apparel retailing. The scope of the current study has been limited to the proof of concept and the future work is planned towards extending the retail analytics to include re-mining as an important component.

Acknowledgement

This work is financially supported by the Turkish Scientific Research Council under Grant TUBITAK 107M257.

References

1. Brijs, T., Swinnen, G., Vanhoof, K., Wets, G.: Building an association rules framework to improve product assortment decisions. Data Min. Knowl. Discov. 8(1), 7–23 (2004)
2. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: Buneman, P., Jajodia, S. (eds.) SIGMOD Conference, pp. 207–216. ACM Press, New York (1993)
3. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In: Jagadish, H.V., Mumick, I.S. (eds.) SIGMOD Conference, pp. 1–12. ACM Press, New York (1996)
4. Han, J., Kamber, M.: Data Mining Concepts and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2006)
5. Korn, F., Labrinidis, A., Kotidis, Y., Faloutsos, C.: Quantifiable data mining using ratio rules. VLDB J. 8(3-4), 254–266 (2000)
6. Angiulli, F., Ianni, G., Palopoli, L.: On the complexity of inducing categorical and quantitative association rules. Theoretical Computer Science 314(1-2), 217–249 (2004)
7. Ertek, G., Demiriz, A.: A framework for visualizing association mining results. In: Levi, A., Savaş, E., Yenigün, H., Balcısoy, S., Saygın, Y. (eds.) ISCIS 2006. LNCS, vol. 4263, pp. 593–602. Springer, Heidelberg (2006)
8. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: An overview. In: Advances in Knowledge Discovery and Data Mining, pp. 1–34. AAAI Press, Menlo Park (1996)
9. Savasere, A., Omiecinski, E., Navathe, S.: Mining for strong negative associations in a large database of customer transactions. In: Proceedings of the 14th International Conference on Data Engineering, pp. 494–502 (1998)
10. Tan, P.N., Kumar, V., Kuno, H.: Using sas for mining indirect associations in data. In: Western Users of SAS Software Conference (2001)
11. Aumann, Y., Lindell, Y.: A statistical theory for quantitative association rules. J. Intell. Inf. Syst. 20(3), 255–283 (2003)
12. Yao, Y., Zhao, Y., Maguire, R.B.: Explanation-oriented association mining using a combination of unsupervised and supervised learning algorithms. In: Xiang, Y., Chaib-draa, B. (eds.) Canadian AI 2003. LNCS (LNAI), vol. 2671, pp. 527–531. Springer, Heidelberg (2003)
13. Yao, Y., Zhao, Y.: Explanation-oriented data mining. In: Wang, J. (ed.) Encyclopedia of Data Warehousing and Mining. Idea Group Inc., USA (2005)
14. Antonie, M.L., Zaïane, O.R.: An associative classifier based on positive and negative rules. In: Das, G., Liu, B., Yu, P.S. (eds.) DMKD, pp. 64–69. ACM, New York (2004)
15. Ng, R.T., Lakshmanan, L.V.S., Han, J., Pang, A.: Exploratory mining and pruning optimizations of constrained associations rules. In: SIGMOD 1998: Proceedings of the 1998 ACM SIGMOD international conference on Management of data, pp. 13–24. ACM, New York (1998)