A Framework for Automated Association Mining Over Multiple Databases

Esma Nur Çinicioğlu
Faculty of Business Administration
Istanbul University
Istanbul, Turkey

Gürdal Ertek
Faculty of Engineering and Natural Sciences
Sabanci University
Istanbul, Turkey

Deniz Demirer
Faculty of Engineering and Natural Sciences
Sabanci University
Istanbul, Turkey

Hasan Ersin Yörük
Faculty of Engineering and Natural Sciences
Sabanci University
Istanbul, Turkey

Abstract—**Literature on association mining, the data mining methodology that investigates associations between items, has primarily focused on efficiently mining larger databases. The motivation for association mining is to use the rules obtained from historical data to influence future transactions. However, associations in transactional processes change significantly over time, implying that rules extracted for a given time interval may not be applicable for a later time interval. Hence, an analysis framework is necessary to identify how associations change over time. This paper presents such a framework, reports the implementation of the framework as a tool, and demonstrates the applicability of and the necessity for the framework through a case study in the domain of finance.**

Keywords- ***association mining over multiple databases; association mining; data mining; association mining visualization; graph visualization***

I. INTRODUCTION

Association mining is a popular framework within data mining, and investigates the association relationships between the items in transactions and attributes in data. Association mining produces interpretable and actionable results, in the form of itemsets or rules, with computed values of interestingness metrics, such as support and confidence. The methodology is used increasingly among practitioners and business analysts [1].

From a practical point of view, the methodology and application literatures have two important gaps that need to be filled: Firstly, interpretation of association mining results that facilitate policy making, Secondly, the automatic execution of association mining for multiple subsets of the same database, such as transactions in multiple time periods, and comparative analysis of the results for these multiple databases.

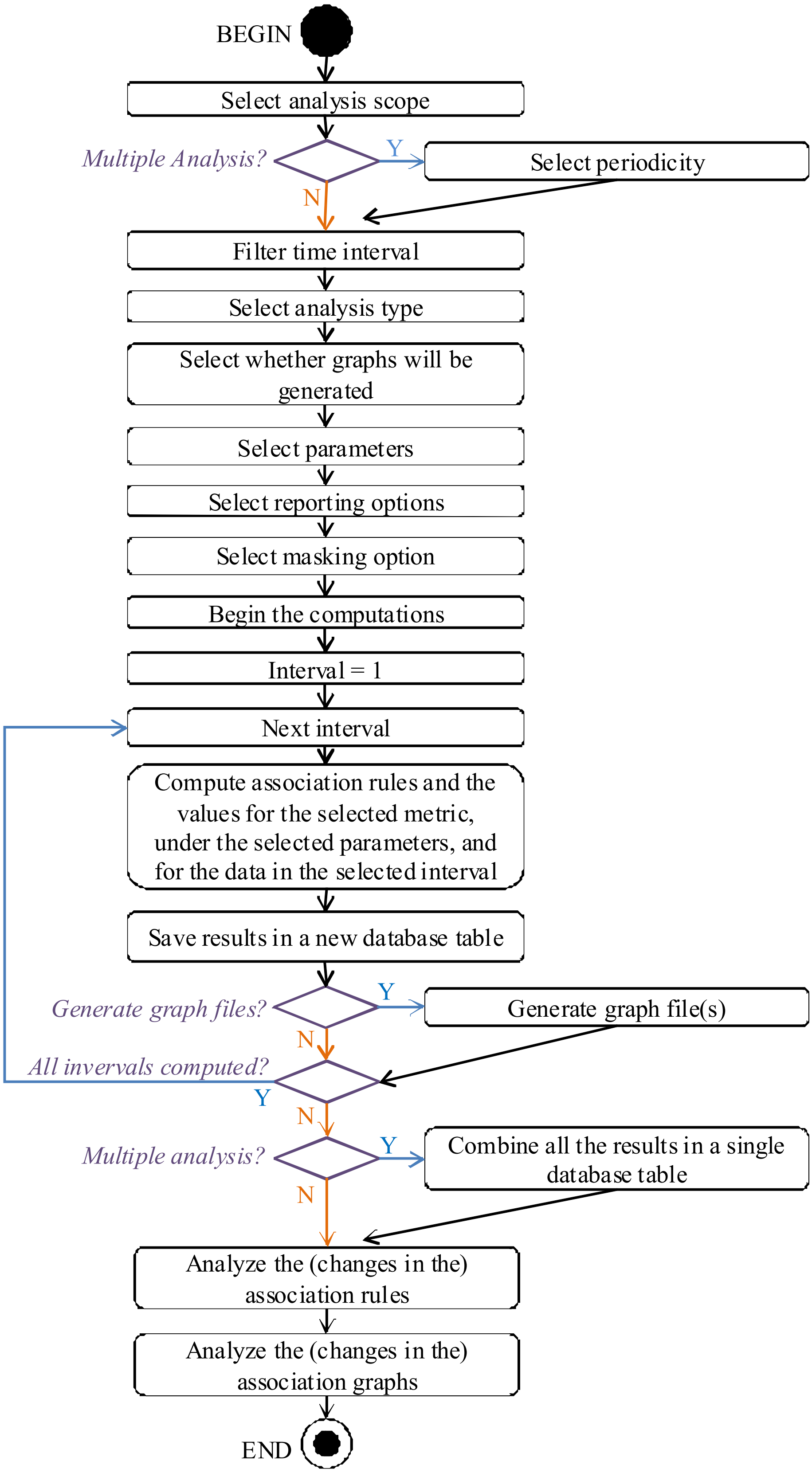


Figure 1. The proposed framework for multiple association analysis.

The simultaneous need for the two mentioned issues was first observed in a consulting project of the first author in 2005, in analyzing automotive spare parts sales. Under the author's supervision, one of the largest importers of automobiles in the country implemented an expert system for the cross-selling of spare parts for pickup vehicles. The rules of the expert system were based on association rules derived from a long, tedious and overwhelmingly manual association mining process. The author and his research team observed great changes in association patterns, even for transactions spaced two months apart.

In this paper, three main contributions are made, covering both methodology and application:

1) A framework is developed to carry out association mining over multiple databases and merge the results in a structured way, enabling posterior analysis. The process through which data from multiple association mining sessions can be combined and the supporting database schema is described throughout the paper.

2) As a practical contribution, and research and analysis facilitator, the AssocMiner software has been developed as an implementation of the framework.

3) The applicability and usefulness of the framework is demonstrated through a case study, where association mining results for financial data is visually analyzed. It is shown that significant differences exist between the association graphs of multiple databases that are temporally separated.

The framework presented in this paper accomodates a multitude of types of association mining analysis, including frequent itemsets, closed itemsets, maximal itemsets, association hyperedges and association rules. In the case study of the paper, frequent itemsets, based on single-dimensional, single-level boolean association rules [2] are analyzed and discussed.

II. LITERATURE

The basic aspect of data mining is finding associations between the entities that provide meaningful relationships. However, one should remember that the essential goal is not only finding associations, but to come up with associations that reveal the hidden information inherent in the database. However, most of the transactional databases -and the patterns that they contain- evolve through time, which makes it impossible and also unreliable to comprehend with a single evaluation. Therefore it is necessary to perform the analysis through a temporally adaptive framework, allowing the user to select the time intervals for which the data should be evaluated. Reference [3] proposes a user-centric rule filtering method, that allows identifying association rules that exhibit a certain user-specified temporal behavior with respect to rule evaluation measures. The approach in [3] is built on the requirement that temporally ordered sets of association rules are available.

A large set of rules can be generated using association rule mining. But the real task is to identify those rules that generate the most interesting insights for the user. With that purpose, [4] defines a new optimized rule mining problem that allows a partial order in place of the typical total order on rules. By solving this optimized rule mining problem with respect to a particular partial order, most-interesting rules can be identified, based on several interestingness metrics. Reference [5] suggests the use of rule templates to describe the structure of interesting rules.

In addition to data-driven interestingness measures, there also exist the claim by many researchers that the most interesting rules can only be found with the help of user integration and user knowledge in the search process. With that in mind, [6] integrates the user and the user knowledge in the filtering and pruning tasks, by representing user knowledge using ontologies. Reference [7] proposes an approach to assist the user in finding interesting rules from a set of discovered association rules. In a later work, [8] suggests that a rule is only interesting in the context of other rules, and the main problem is that most of the existing methods treat rules individually. As a solution, [8] proposes that a major part of rule exploration can be handled as an On-line Analytical Processing problem.

While the detection and selection of the hidden information inside a database are two essential goals for association rule mining, most of the time it is the delivery method that provides an understandable platform to users. To this end, visualization plays an important role in association mining. Reference [9] proposes a novel approach where the user is enabled to drive his/her navigation through the voluminous rule set by trial and error via the successive limited subsets s/he focuses on. Visualization helps the user in comprehending the bulk of rules and finding the ones that are most relevant for decision-making. An important aspect of successful visualization in data mining is the simultaneous achievement of both the global view and the details of the database. For that purpose, different approaches are developed which enable the display of large databases with large sets of association rules [10], [11]. There exists an extensive literature for the visualization of association rules [12] [13] [14]. Reference [15] claims that data and rule visualizations should be integrated, facilitating an improved understanding of rules. Reference [16] develops a frequent itemset visualizer to display mined frequent items. The visualization of data mining results is important in many different domains, including the visualization of web information. A visualization tool to visualize web graphs is proposed in [17]. With the web graph algebra proposed in their work, web graphs and their layers can be combined and manipulated to discover new patterns. An extensive survey of the state of art application of visualization techniques in terms of visualization of derived rules, visualization of rules, and visual interactive rule derivation can be found in [18].

III. THE PROPOSED FRAMEWORK

The framework that is proposed for multiple association mining analysis is presented in Figure 1, and is described in this section.

A. *Analysis Process*

The analysis process in our proposed framework is given in Figure 1. The analysis begins with the selection of the scope of the analysis, where the analyst can select either single analysis, or automated multiple analysis. For analysis over multiple databases, periodicity is selected. The time interval for which the analysis will be performed, together with the type of analysis are selected next. Once the parameters, reporting options and masking option are selected, the computations are initiated. For each time interval, a subset database is constructed, and computations are performed. The results are saved in separate files following the database structure in Tables 1-3, and association graphs are generated for each of the subset databases. The results for the multiple analyses are then combined in a single database for integrated analysis of changes in associations.

B. *Database Schema*

The database schema for the inputs and outputs for the presented framework are given in Tables 1, 2, and 3.

Table 1 shows the table structure for the input transactions database. The first field in the transactions database is *Date*. If the software is to be run for a single analysis, and there is no date information available, the modeler can simply create a fictitious date and fill the first column with that value.

TABLE I. TABLE STRUCTURE FOR INPUT TRANSACTIONS DATA

Date
TransactionID
ItemID

TABLE II. TABLE STRUCTURE FOR MAXIMAL ITEMSETS, CLOSED ITEMSETS, FREQUENT ITEMSETS, AND ASSOCIATION HYPEREDGES

Start_Date[a]
AssocID
Abs_Support[b]
Support
Confidence[c]
Item1
...
ItemM

a. Only for the combined association rule database in the case of multiple databases.

b. Optional.

c. Only for association hyperedges.

Table 2 shows the table structure for the outputs of maximal itemset, closed itemsets, frequent itemsets and association hyperedges analyses. These analyses generate itemsets in the form of {A, B}, rather than rules of the form A =>B. So there is no consequent in these types of analyses, only itemsets. In case of association mining over multiple databases, the first field, *Start_Date*, denotes the time interval that the itemset belongs. The third field, *Abs_Support*, shows absolute support (support count, the number of transactions that contain the itemset), and is optional. The fifth field, *Confidence*, is applicable only if the analysis type is association hyperedges.

TABLE III. TABLE STRUCTURE FOR ASSOCIATION RULES

Start_Date[a]
AssocID
Abs_Support[b]
Support
Confidence
Conseq
Antec1
...
AntecM

a. Only for the combined association rule database in the case of multiple databases.

b. Optional.

Table 3 shows the table structure for the output of association rule analysis. Unlike the other analyses, association rule analysis results in rules of the form A=>B, where A is the set of items in the antecedent, and B is the set of items in the consequent. The association rules in this study were restricted to those with a single item in the consequent, and this item is given under the column *Conseq*, before all the antecedents (*Antec1*, ..., *AntecM*).

The primary key in Table 1 is a composite key (a key that consists of two or more attributes), consisting of all the three attributes. The primary key in Tables 2 and 3 depend on the analysis scope. Of the analysis performed over multiple databases, the primary key is a composite key that consists of *Start_Date* and *AssocID*. Alternatively, for single analysis, the primary key is *AssocID*.

C. Graph Visualizations

The generation and analysis of association graphs is an integral part of the framework. The visualization scheme selected for the framework is the scheme introduced by [1], that visualizes the results of the association mining as a directed graph. In the visualizations, the items, the itemsets, and the association rules are all represented as nodes. Edges represent the links between the items and the itemsets or associations.

In visualizing frequent itemsets, the nodes that represent the items are shown with a thin line, whereas the nodes that represent the itemsets have thicker lines, with the thickness increasing with the cardinality of the itemsets. The sizes (the areas) of the nodes show the support levels. The directed edges symbolize which items constitute a given frequent itemset. The main idea in the adopted framework is to exploit already existing graph drawing algorithms [19] and the software in the information visualization literature [20] for visualization of association mining results which are generated by existing algorithms (and software) in the data mining literature [2].

IV. IMPLEMENTATION OF THE FRAMEWORK: ASSOCMINER

The described framework has been implemented as a software system, called AssocMiner, and has been tested with multiple transactions databases. While the technology-related aspects of the software are explained in the Appendix, the analysis process in AssocMiner is explained here.

A. Selecting the Analysis Scope and Periodicity

AssocMiner begins with a selection of the scope of the analysis, where the analyst can select either single analysis, or automated multiple analysis (Figure 2). In single analysis, only a single run of association mining computations is carried out, and the complete database can be mined. In multiple analysis, subsets of the transactions database are subsequently mined in an automated fashion, without manual intervention during the computations.

Figure 2. Selecting analysis scope.

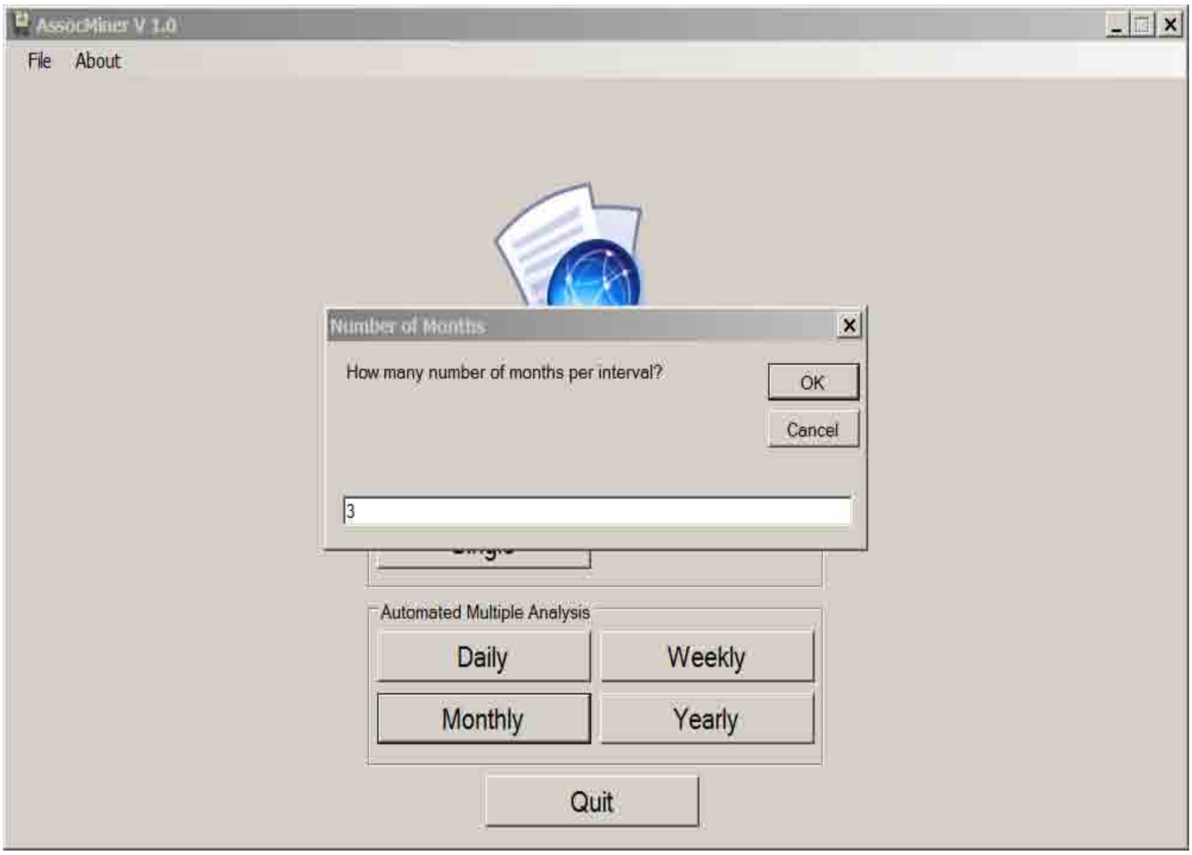
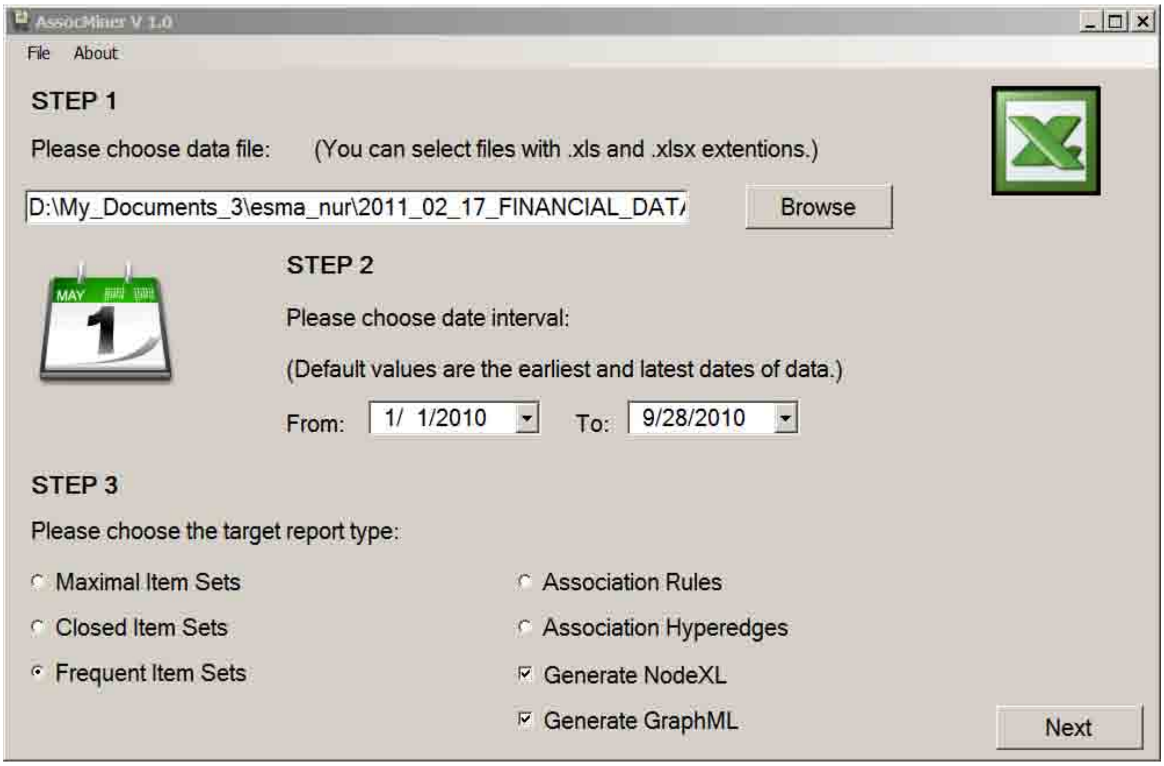


Figure 3. Selecting periodicity.

The multiple databases (subsets) of the original full transactions database are constructed based on time intervals. The time interval for each subset of transactions is a multiple of days, weeks, months, or years. The periodicity is how many periods of the selected time unit will be included in each subset. For example, if an analyst is interested in the analysis of quarterly changes in the associations, s/he should select months in the main window, and then specify 3 in the text box as the number of months in each time period (Figure 3).



Filtering the time frame and selecting the association mining analysis.

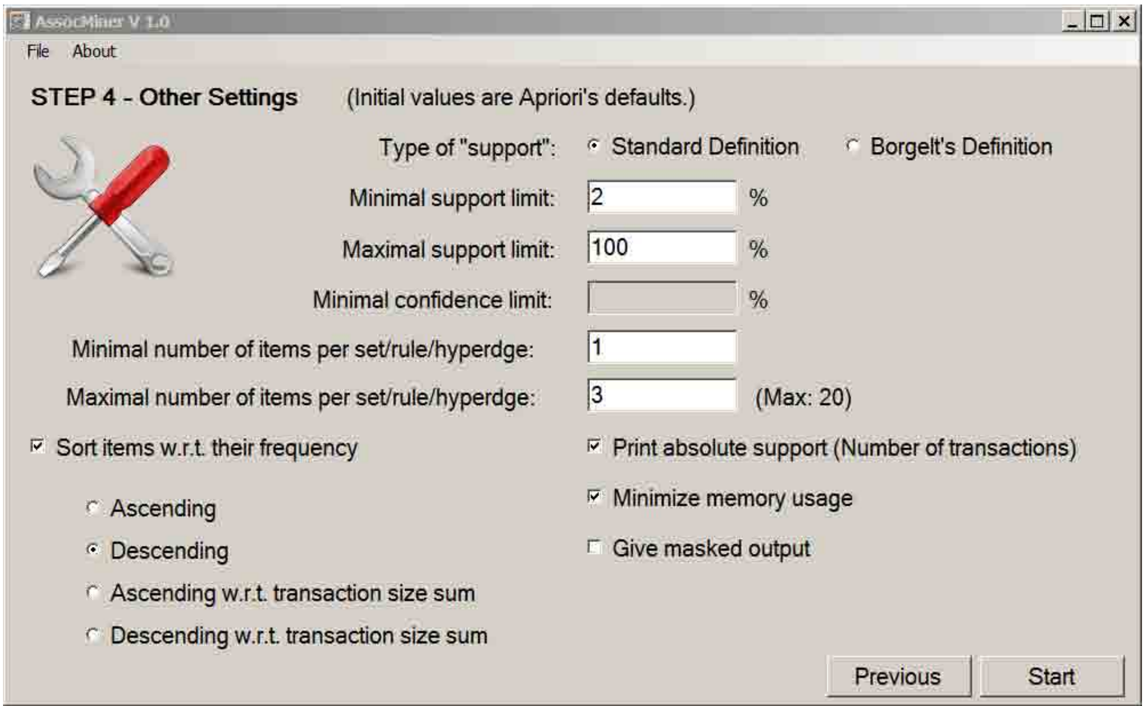


Figure 4. Selecting the parameters, output format, and masking options.

B. *Filtering the Time Frame and Selecting the Association Mining Analysis*

Once the scope of analysis is selected, then the next step is initiated (Figure 4), where the user selects the transactions database file (a MS Excel xlsx spreadsheet file), filters the time interval that s/he is interested in, selects the type of association mining to be conducted, and specifies whether s/he would like to obtain the graph visualization file. Graph visualization is generated as a graphml file, which can subsequently be processed in yEd graph visualization software[1]. Graph visualization is currently implemented only for analyses that generate itemsets (all analyses except association rule analysis).

C. *Selecting the Parameters, Output Format, and Masking Options*

Once the type of association mining is selected, the analyst enters the parameters for the Apriori algorithm. Minimal/maximal support, minimum/maximum number of items are specified for all types of association mining (Figure 5). Minimal confidence is selected only for association rules and association hyperedges. At this point, the analyst can also specify whether s/he wants to have the itemsets/rules sorted with respect to their frequencies (absolute support values), have the absolute support printed, and minimize memory usage. One critical feature of AssocMiner is that it can generate masked outputs, hiding the original item and order labels when generating the association mining results. In the case of masking, the analyst can share the anonymous results with others, distributing the visual data mining process to a team of analysts. Look-up files list the original item and order labels corresponding to each of the masked item and order labels.

D. *Running the Analysis and Obtaining the Results*

AssocMiner conducts the association mining by running the Apriori algorithm (Figure 6), and generates the results in the form of MS Excel xlsx and yEd graphml file(s) (Figures 8-10). Each set/rule is printed on a row, and the items in each set/rule are sorted lexicographically. Each set/rule is labeled with a unique label that is constructed through concatenating the strings of the item names.

[1] http://www.yworks.com

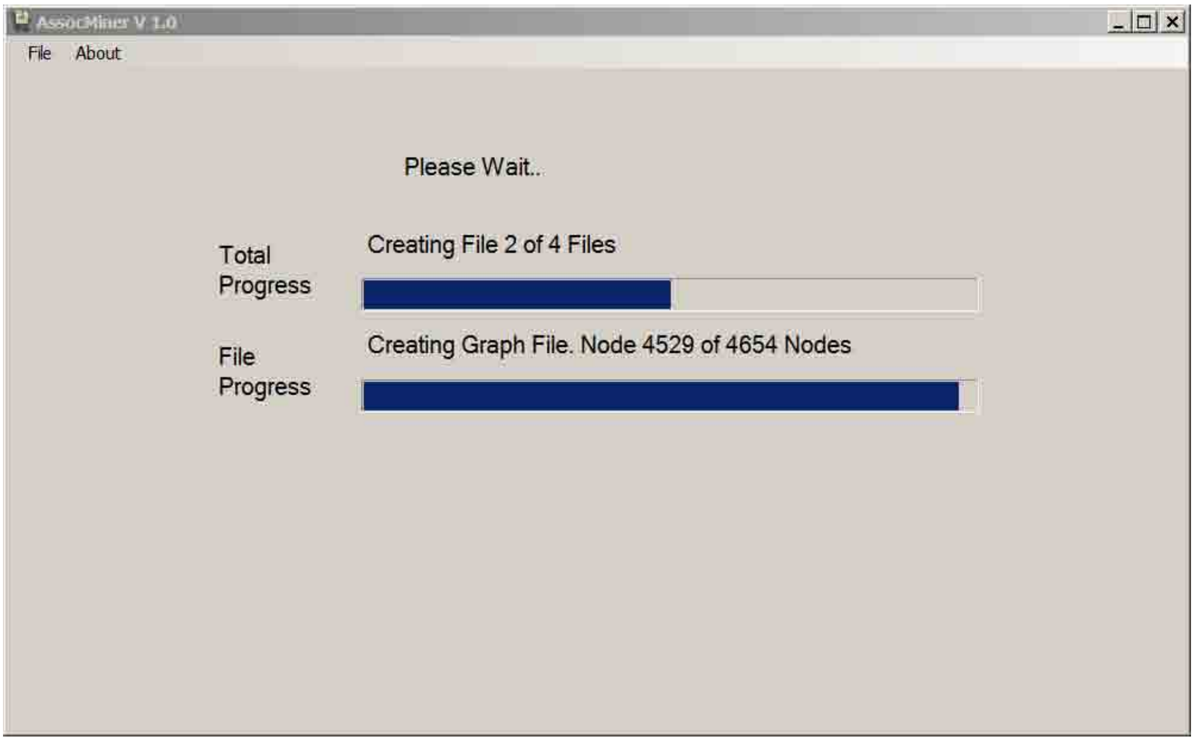


Figure 5. Running of the multiple association mining analysis.

V. CASE STUDY

The applicability of and the necessity for the framework is demonstrated through a case study in this section. The main result of the case study is that associations change significantly over even brief time periods.

A. *The Data*

Finance is a popular domain where multi-variate temporal data is abundantly available and is extensively analyzed. To this end, data was collected on the exchange rates for Dollar, Euro, and Gold. Then the original time-stamped numerical database was transformed into a time-stamped Market Basket Analysis (MBA) database, where each day is considered as a distinct transaction. The market basket for each day can include 12 types of items: The discretized daily change in prices of each of the three commodities for today, yesterday, the day before, and three days ago.

For example, TransactionID=40180 for the day Jan 2, 2010 consists of the following items:

{DollarDecr03Today, DollarDecr05DayAgo1, DollarDecr03DayAgo2, DollarDecr06DayAgo3, EuroDecr01Today, EuroDecr03DayAgo1, EuroDecr07DayAgo2, EuroDecr07DayAgo3, GoldDecr01Today, GoldDecr03DayAgo1, GoldDecr05DayAgo2, GoldDecr07DayAgo3}

This means that the purchase price of dollar decreased [0.3, 0.4) percent on January 2, compared to January 1, the price of dollar had decreased [0.5, 0.6) percent on January 1, compared to December 31, 2009, etc.

The obtained transaction database consists of 3252 transactions rows, corresponding to the 271 days from Jan 1 to September 28 in 2010. This data was then divided into three databases, covering three time intervals of 3 months each (hereon referred to as intervals 1, 2, and 3).

Figure 6. Change in frequent itemset support values..

B. Results

For visually understanding the change in the frequent itemsets, a treemap visualization [21] is created (Figure 7). In this visualization, the frequent itemsets (those with at least 10% support) are shown with labeled grey containers, the support values of frequent itemsets in each time interval are reflected in the sizes of colored rectangles, and the time interval is shown with color (increasing darkness as for higher time interval values). If the support values for itemsets were the same over the intervals, than one would observe three equally sized colored rectangles in each container. However, this is not the case for most of the containers (itemsets). This visualization shows that frequent itemsets change significantly over time, even in intervals of 3 months length.

Numerical values for the number of frequent itemsets in each interval give further insights: Total of 177 itemsets that have support values greater than 10% in any of the three intervals. 15 of them have appeared (had support of at least 10%) in all three intervals, 33 of the rules have appeared in only two intervals, and 129 of the rules have appeared in only one interval.

Statistical analysis (Friedman's nonparametric test [22] for differences in means) suggests that there is statistically significant difference between the support values of the 177 itemsets in the three intervals. It is striking that only 8 percent of associations (15/177=0.0847) are observed in all the three intervals. This brings out the deeper and extremely important question of how applicable association mining really is, which rises as a result of the case study.

Figures 8, 9, 10 show the association graph for the three time intervals for which association mining is carried out. Even a quick visual observation clearly shows that the three graphs are structurally different: The association graph for interval 1 (Figure 8) has many itemset (lined up at the bottom of the figure) that are not associated with other nodes, whereas almost all the remaining ones are tightly linked. For interval 2 (Figure 9), the number of associations is less, and the intensity of the links is weaker. For interval 3 (Figure 10), the clustering pattern is similar to that of interval 1, but the number of independent itemsets (at the bottom) is much fewer.

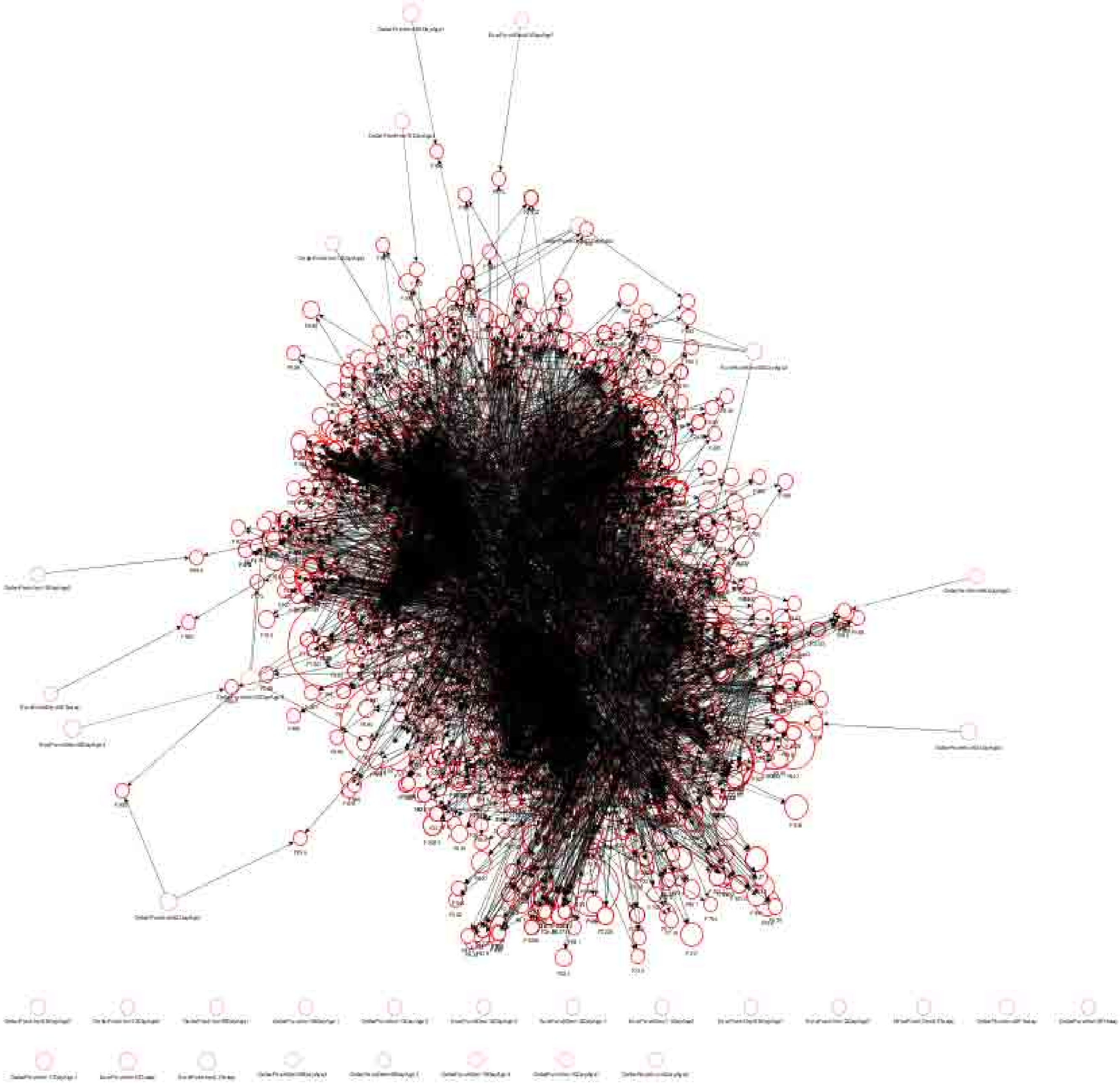


Figure 7. Association graph for time interval 1.

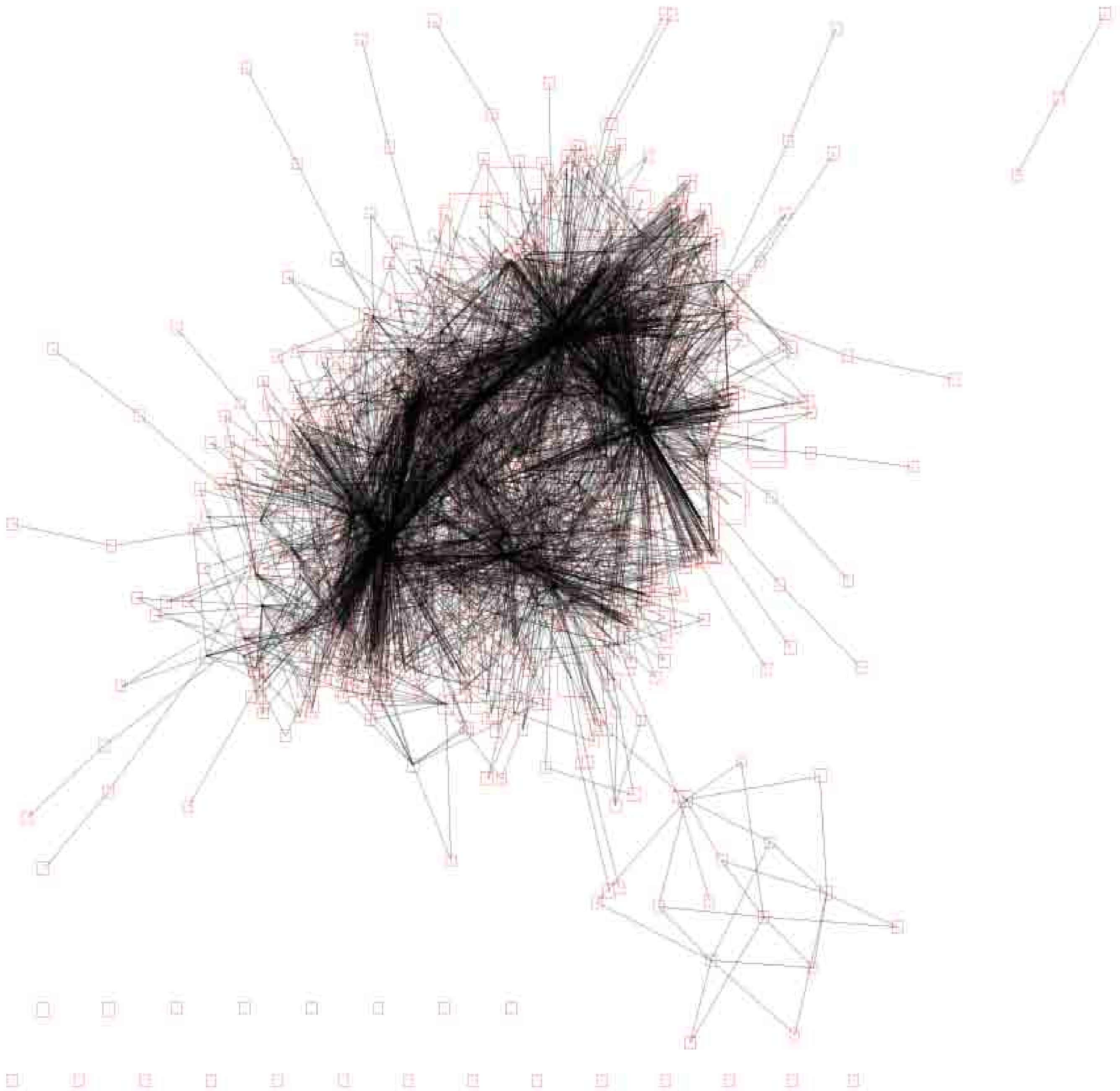


Figure 8. Association graph for time interval 2.

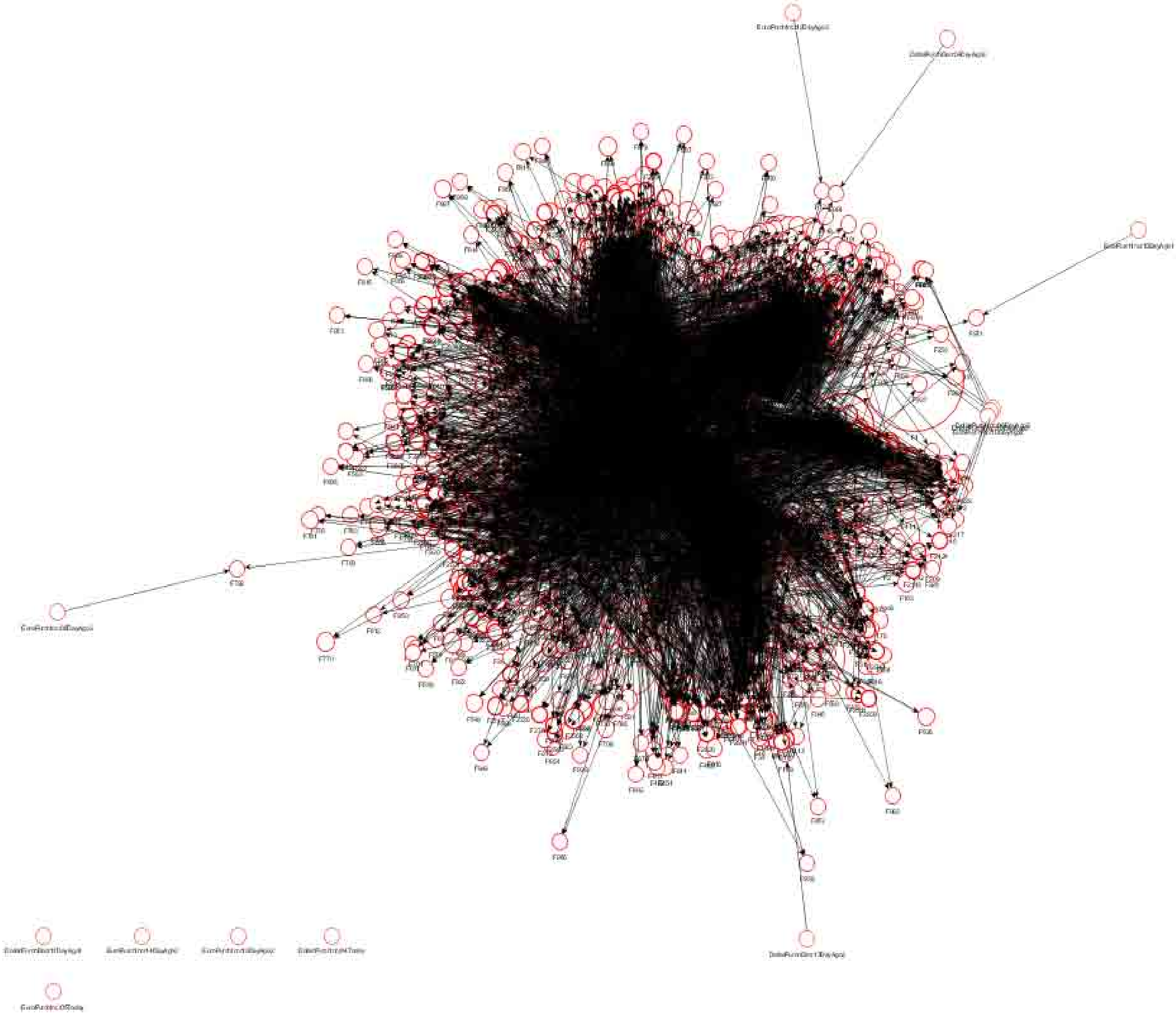


Figure 9. Association graph for time interval 3.

VI. CONCLUSIONS AND FUTURE WORK

The analysis results in the case study prove that significant changes can occur in association patterns in successive time intervals. Hence the multiple temporally successive databases that represent the transactions have to be mined systematically through a convenient methodology. This temporal change in associations was the major motivation of our study, and was first observed by the first author in a consulting project in 2005, in analyzing automotive spare parts sales. The development of the presented framework -and the AssocMiner software that implements the framework- enabled the observation of the same observation for the financial database of the case study. As a future research, similar studies can be conducted in other domains where temporal changes in transactional patterns are of interest, and observe whether the same phenomena exists in those domains.

ACKNOWLEDGMENT

The authors thank Ayhan Demiriz, Enes Eryarsoy, and Fatih Çakmak for sharing their knowledge and experience on the topic, and Armağan Aktuğ and Mehmet Arabacı for collecting and preparing the data for the case study. The authors also thank Mehmet Can Arslan for long and deep discussions on the change in association patterns over time, which motivated this research. Finally, the authors thank Tilanka Kulatunga for his help in completing the software development.

APPENDIX: TECHNOLOGY SELECTION AND TOOLS USED

The MS Windows[2] operating system (OS) was selected for the AssocMiner implementation, since this OS is the de facto standard in the global business community.

Visual Basic (VB)[3] has been selected as the programming language, since it is the default programming language for MS Excel[4]. MS Excel is the leading commercial spreadsheet software throughout the world, and is extremely popular in business applications for data analysis purposes. Once VB was selected, MS Visual Studio.NET[5] was the natural selection as the IDE (integrated development environment).

Christian Borgelt's Apriori program[6], which is coded in the C language, was used as the association mining engine in the project. AssocMiner externally calls Borgelt's executable command-line Apriori program, and then parses the text results into MS Excel xlsx files and yEd graphml files. This program was selected because it has been coded efficiently and runs very fast even for large input databases. The source code for this program has been used earlier as the

[2] http://www.microsoft.com/WINDOWS/
[3] http://msdn.microsoft.com/en-us/vbasic/default
[4] http://office.microsoft.com/en-us/excel/
[5] http://msdn.microsoft.com/en-us/vstudio/default
[6] http://www.borgelt.net/apriori.html

association mining computational engine in commercial SPSS Clementine software (now IBM SPSS Modeler[7]), further motivating its selection.

The associations are represented as graphs (except association rules), and a yEd graphml file is generated if the analyst asks for an association graph. The graphml file format is based on XML[8], the standard markup language for flexible specification and exchange of data in the internet-age. The resulting graphml files are opened in the yEd software[9], a very advanced graph drawing software, which has an intuitive user interface.

In the case study section, Treemap visualization of Figure 7 was created using Visokio Omniscope[10] software. Statistical testing of the difference between support values was conducted in SPSS[11] statistical software, using nonparametric Friedman's test [22]. Graph visualizations in Figures 8-10 were created using yEd software.

References

[1] G. Ertek and A. Demiriz, "A framework for visualizing association mining results," Lecture Notes in Computer Science, vol. 4623/2006, Springer Berlin / Heidelberg, 2006, pp. 593–602.

[2] J. Han and M. Kamber, "Data mining: concepts and techniques," Morgan Kaufman Publishers, 2001.

[7] http://www.spss.com/software/modeler/
[8] http://www.w3.org/standards/xml/
[9] http://www.yworks.com/
[10] http://visokio.com/
[11] http://www.spss.com/

[3] M. Steinbrecher and R. Kruse, “Visualizing and fuzzy filtering for discovering temporal trajectories of association rules,” Journal of Computer and System Sciences, February 2010, 76(1), pp.77–87.

[4] R. J. Bayardo, Jr. and R. Agrawal, “Mining the most interesting rules,” in Proc. 5th Int. ACM SIGKDD Conf. Knowledge Discovery Data Mining, 1999, pp. 145–154.

[5] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A.I. Verkamo, “Finding interesting rules from large sets of discovered association rules,” In Proc. 3rd Int’l Conf. on Information and Knowledge Management, November 1994, Gaithersburg, Maryland, pp. 401–408.

[6] C. Marinica, F. Guillet, and H. Briand, “Post-processing of discovered association rules using ontologies,” The Second International Workshop on Domain Driven Data Mining, Pisa, Italy, 2008, pp. 126–133.

[7] B. Liu, W. Hsu, S. Chen, and Y. Ma, “Analyzing the subjective interestingness of association rules,” IEEE Intelligent Systems 15(5), 2000, pp. 47–55.

[8] B. Liu, K. Zhao, J., Benkler, and W. Xiao, “Rule interestingness analysis using OLAP operations,” In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (Philadelphia, Aug. 20–23). ACM Press, New York, 2006, pp. 297–306.

[9] J. Blanchard and F. Guillet, “Exploratory visualization for association rule rummaging,” In 4th International Workshop on Multimedia Data Min-ing MDM'03 in con-junction with KDD'03, 2003, pp. 107–114.

[10] O. Couturier, J. Rouillard, and V. Chevrin, “An interactive approach to display large sets of association rules,” In Proceedings of the 12th International Conference on Human-Computer Interaction (HCI’07), Beijing, China, 2007.

[11] O. Couturier, T. Hamrouni, S. Ben Yahia, and E. Mephu Nguifo, “A scalable association rule visualization towards displaying large amounts of knowledge,” In Proceedings of 11th International conference on Information Visualization IV’07, Zurich, Switzerland, IEEE Computer Society, July 2007, pp. 657–663.

[12] P. Buono and M. F. Costabile. "Visualizing association rules in a framework for visual data mining," In M. Hemmje et al., editor, From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments, Springer-Verlag, February 2004, pp. 221–231.

[13] S. Chakravarthy and H. Zhang, "Visualization of association rules over relational DBMSs," In Proc. 2003 ACM Symp. on Applied Computing, ACM Press, Melbourne, Florida, 2003, pp. 922-926.

[14] T. Herawan, I. T. R. Yanto, and M. M. Deris, "SMARViz: Soft maximal association rules visualization," Lecture Notes in Computer Science, 2009, vol. 5857/2009, pp. 664-674

[15] Y. Liu and G. Salvendy, "Design and evaluation of visualization support to facilitate association rules modeling," International Journal of Human-Computer Interaction, 27(1), 2006, pp.15-38.

[16] C.K.-S. Leung, P.P. Irani, and C.L. Carmichael, "WiFIs-Viz: Effective visualization of frequent itemsets," InProc. IEEE ICDM 2008, pp. 875–880.

[17] J. Chen, L. Sun, O. R. Zaiane and R. Goebel, "Visualizing and discovering web navigational patterns," In Proceedings of the 7th International Workshop on the Web and Databases: collocated with ACM SIGMOD/PODS, Paris, France, 2004, pp. 13–18.

[18] Y. Liu and G. Salvendy, "Visualization to facilitate association rules modelling: A review," Ergonomia IJE&HF, 2005, vol. 27, No. 1, pp.11–23.

[19] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis, "Graph drawing: algoritms for the visialization of graphs," Prentice Hall PTR, 1998.

[20] I. Herman, G. Melançon, M. S. Marshall, "Graph visualization and navigation in information visualization: A survey," IEEE Transactions on Visualization and Computer Graphics. 6 No.1,2000, pp. 24–43.

[21] B. Shneiderman, "Tree visualization with tree-maps: 2-d space-filling approach," ACM Transactions on Graphics, vol. 11(1), Jan. 1992, pp. 92-99

[22] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," Journal of the American Statistical Association, Vol. 32, No. 200 ,Dec. 1937, pp. 675-701.