A Theoretical Model for Pattern Extraction in Large Datasets

MUHAMMAD USMAN*, MUHAMMAD AKRAM SHAIKH

Pakistan Scientific and Technological Information Center, Islamabad, Pakistan

**Corresponding author's e-mail : usmiusman@gmail.com*

Abstract

Pattern extraction has been done in past to extract hidden and interesting patterns from large datasets. Recently, advancements are being made in these techniques by providing the ability of multi-level mining, effective dimension reduction, advanced evaluation and visualization support. This paper focuses on reviewing the current techniques in literature on the basis of these parameters. Literature review suggests that most of the techniques which provide multi-level mining and dimension reduction, do not handle mixed-type data during the process. Patterns are not extracted using advanced algorithms for large datasets. Moreover, the evaluation of patterns is not done using advanced measures which are suited for high-dimensional data. Techniques which provide visualization support are unable to handle large number of rules in a small space. We present a theoretical model to handle these issues. The implementation of the model is beyond the scope of this paper.

Keywords: Association Rule Mining, Data Mining, Data Warehouses, Visualization of Association Rules

INTRODUCTION

Knowledge discovery in databases (KDD) has been gaining popularity in the recent past. Now-a-days techniques have been developed to extract knowledge from large data sets, where data is in static form as well as in the shape of streams [1]–[6]. Pattern extraction has been applied at large datasets in past to extract frequent patterns from a dataset. Such patterns are not only hidden but are also interesting to the business analysts. There has been a recent trend of strengthening the pattern extraction techniques by deploying techniques of multi-level mining, dimension reduction, advanced evaluation and visualization support [3].

Modern techniques of pattern extraction work at multiple levels of abstraction which allows the capability to mine at different levels in the hierarchy rather than the whole dataset at once. Methods including Agglomerative Hierarchical Clustering, K-Means Clustering *etc.* have been used in this regard. However these methods are not effective for mixture of nominal and numeric data. Furthermore, these techniques do not handle large number of dimensions effectively. Authors have also emphasized in past to deploy dimension reduction techniques to pick top ranked dimensions for ranking process. However, there are flaws in existing dimension reduction approaches. These approaches work for nominal and numeric variables separately. Moreover, these approaches don't consider the semantic relationship between these attributes. Another issue is the usage of *Apriori* algorithm which doesn't suite for large size datasets. Pattern extraction, now-a-days, is being done on aggregate data, for example, a data warehouse. In such cases data gets larger in volume as well as the number of dimensions increase. The authors in past have used conventional measures to evaluate the interestingness of the extracted patterns in general. Some authors have used advanced measures of interestingness like Lift, Loevinger, Rae, CON and Hill [2], [7], [8]. If patterns are extracted from aggregate data, measures which suite for aggregate data should be used for evaluation of interestingness. Another key aspect of the mining process is its ability to visualize patterns using a graphical interface. There are techniques which provide visualization support for extracted patterns using Bar Graphs, Ball Graphs [1], or Semiology Principle [7]. However, these techniques are unable to show large number of rules in a small graphics interface as luminosity and color shades affect these approaches. An effective approach is required where the expert can interact with the extracted patterns graphically to explore patterns by comparing their interestingness parameters within the interface.

In this paper, we have reviewed different techniques which provide robust methods for pattern extraction in large datasets. It is evident that there is a need of a robust methodology to mine at multiple levels which can be effective for large datasets. Moreover, this methodology needs to handle both nominal and numeric data in order to create levels in the hierarchy. Secondly, in the available literature, dimension reduction is done separately for nominal and numeric attributes. The dimension reduction techniques need to employ a single method for ranking dimensions for nominal and numeric data. In most cases, conventional algorithm (*Apriori*) is used for patter extraction. It is proposed that the modified algorithm for pattern extraction in large datasets is used instead. If mining is being applied on aggregate data, measures which suite for aggregate data, should be used for evaluation of interestingness. In some cases, authors have provided visualization support but these approaches don't handle large number of rules effectively. The interface should be able to present large number of rules efficiently using different interestingness parameters.

In order to overcome the limitations of the existing approaches, we propose a model which can be used to extract interesting patterns from large datasets. In the proposed model, effective techniques like multi-level mining, dimension reduction, advanced evaluation and visualization are recommended.

LITERATURE REVIEW

This Section presents the previous work done in the area of pattern extraction in large datasets.

Data mining in large datasets has been discussed by [3] in detail. The authors emphasize on the usage of multi-level mining, ability to mine without domain knowledge expertise, usage of advanced measures of interestingness and visualization ability while mining in large datasets.

On similar lines, [9] discuss the challenges of recent data mining techniques in recent days. Authors accentuate that such techniques should be seen in terms of their ability to handle the large size databases and high dimensionality as well as the ability to work with complex relationships within the dataset between different variables. Authors also suggest that the business analysts may be given the choice of reducing the number of dimensions in order to remove the irrelevant variables involved in decision making. Moreover, more efficient algorithms handling the large size data can be used.

Generally Hierarchical clustering is used to create clusters at different hierarchical levels. These clusters are further used for mining at each level in the hierarchy. Such mining process provides ability to mine at different levels on small subsets of a dataset rather than the whole dataset. However, this clustering technique only deals with numeric data while creating clusters.

In order to overcome the limitations of Hierarchical Clustering Algorithm, Li and Biswas [10] worked on an extension of Hierarchical Clustering algorithm. The extended algorithm deals both numeric and nominal data, and is called SBAC (Similarity-based Agglomerative Clustering). The approach works with e mixed data measure schema. Moreover, features values with less common matches are given extra consideration.

In another work, [11] conducted a survey of clustering algorithms. Authors provide the comparison on the basis of complexity and capability to handle high dimensional data. Authors suggest that AHC, K-Means and DBSCAN along with few other algorithms don’t have the capability to handle high dimensional data. Moreover, algorithms like CLIQUE, ORCLUS etc can handle high dimensional data more effectively. The comparison is made up by applying these algorithms on datasets including IRIS, Mashrooms, Gene Expression Data and Traveling Salesman Problem etc.

In a different approach, [12] emphasize on creating correlated groups of the whole data set before applying mining techniques on the data. Correlation exists within the data between different variables which enables to create groups of rows. After groups are obtained, association rule mining is applied at group level to obtain association rules against each group. Although such technique is applicable to certain domains, but its efficiency is not known in cases where number of dimensions become higher.

Pattern extraction is usually done by generating association rules from dataset. Each rule presents a pattern which is found to be frequent in the whole dataset. An association rule has two parts. The left part of the pattern has a set of attribute, the occurrence of which predicts the occurrence of the attribute at right side of the rule. In past, different algorithms have been proposed to mine association rules from datasets.

In order to extract patterns in the form of association rules from the dataset Agrawal *et al.* [13] introduced *Apriori* algorithm. This algorithm prunes the item sets by using the closure property in downward direction. This property states that if an item set is not frequent, then its superset is also not frequent. Although anti-monotonicity of item sets is used in the level wise search; the algorithm has to scan the whole database.

In order to improve on some weaknesses of the existing algorithm, Agrawal *et al.*[14] worked on creating an algorithm called *AprioriHybrid*. The proposed algorithm solves the problem of extraction of association rules between items in large datasets. The algorithm is combination of *Apriori* and *AprioriTid* algorithms developed by the authors previously [13]. Authors suggest that the *AprioriHybrid* algorithm which works better than *Apriori* when data set gets larger can be used in situations where data set size increases. However, the implementation of new algorithm is more complex than the previously proposed algorithms. The algorithm uses traditional support measure and is tested on four synthetic data sets.

Yuan [15] developed a new version of well known *Apriori* algorithm. This algorithm focuses on solving two major issues with *Apriori* Algorithms that is, frequent scanning of dataset and large number of candidate sets. In order to avoid scanning of whole datasets, it creates a new mapping way which also improves joining efficiency. An overlap strategy is used to achieve high efficiency of count support. Authors term this algorithm as *T-Apriori Algorithm*. Authors compare the new algorithm with *Apriori*, *I-Apriori* and *Bitxor* algorithms. The authors have used Mushrooms dataset from UCI machine learning repository for implementation purposes. The authors claim the improvement in time consumption upto 98% however the space complexity has not been compared. The algorithm has not been tried on variety of datasets so generalization is not known at this stage.

In another approach, [16] applied association rule mining at cluster level in the dataset to discover interesting links between binary attributes. Rules relevancy has been checked using *Confidence, Lift, OM, Loevinger, ADI* and *Jaccard Indices*. Authors also provided a visualization capability to analyze results generated through this process. The approach has been tested on an industrial database, and has not been generalized. Moreover, the approach only works with binary attributes. It will be interesting to see its application for other types of variables.

In order to extract association rules from datasets, [17] presented an algorithm which works in two passes through the whole dataset. A random sample is selected from the whole data set and it is used to determine the association rules in the dataset. The second pass is then used to see if any rules were missed, and these are added to the resultant set if required. The approach is reported to be effected for large size datasets as it reduces I/O usage to a great deal. However, it has been determined that second pass cannot be avoided at all.

In a different approach, [18] presented a methodology to cluster association rules after extraction. Two methods are proposed by the authors. First method generates a hierarchical structure by creating groups at each level according to structure of rules. The other method measures the semantic distance between rules and groups accordingly. The approach may be effective in places where relevant rules are to be grouped together; however since all association rules must be generated at first, computation time has to be compromised.

In order to advance the rule mining in data warehouses, Zhu [1] worked on association rule mining in multi-dimensional environment. Author divided the multi-dimensional rule mining into different categories. Author defined a category called inter-dimensional association which defines association within a single dimension whereas the intra-dimensional association defines association between multiple dimensions. Author defined hybrid category which combines both types of associations defined before. Author converted data cube to a tabular format to draw frequent items. Author provided visualization support to the mining process using two types of graphs i.e Ball Graph and Bar Graph. However the graphs do not handle the large number of rules in the resultant set effectively.

As a special case, Bogdanova & Georgieva [19] worked on a case study involving *FolkloreCubes*. These cubes are created at Folklore Institute from an archival fund with Folklore material; these data cubes are used to draw patterns in terms of association rules. The dimensional reduction is provided to the user by utilizing the minimum *Support* value. The process has the ability to visualize association rules using a bar graph. Each column in the graph represents a rule. Different heights of columns indicate the value of support of that rule. This visualization is difficult to adapt if large number of rules are to be shown.

In a similar work, Messaoud *et al.* [2] worked on data cubes to mine association rules in a multi-dimensional environment. This approach follows the concept of meta-rules guided mining where meta-rules are created by the user in order to guide the mining process. The discovered rules are interesting for the user as the mining process is guided towards interesting patterns. However hidden interesting patterns are not extracted due to guided process looking for targeted patterns. Evaluation of extracted patterns has been done using *Lift* and *Loevinger* measures. Since the data is available in aggregate form in data cubes, these measures are best suited instead of the conventional measures like *Support* and *Confidence*. The approach is tested using sales data and it was reported to be effective.

The results of the study encouraged the authors for development of a complete environment association rule mining. The new framework included a visualization component. Authors called this framework as the Online Environment for Mining Association Rules (OLEMAR). Associations rules are generated with the help of meta-rule guided mining approach like before and are evaluated using *Lift* and *Loevinger* measures. The graphics *Semiology Principle* is utilized by authors to visualize extracted patterns. Blue squares are used to present the item sets whereas equilateral triangles are used to define the relationship between item sets. The interestingness of rules is shown by using luminosity of the shapes. Usage of luminosity and shape sizes makes it difficult to differentiate different item sets if the results are large in numbers.

In a different approach, Usman *et al*. [8] proposed a technique to generate STAR schema at multiple levels of abstraction in order to mine association rules at different levels. Authors applied agglomerative hierarchical clustering to generate clusters at different level. At each cluster, authors picked top nominal variables using Information Gain which were then used to create STAR Schema. Authors mined rules on STAR schema using *Apriori* Algorithm and evaluated these using advanced measures Rae, CON and Hill. It will be interesting to see if using a different version of hierarchical clustering which handles both numeric and nominal data for clustering, is more effective than this approach. Moreover, for datasets with larger sizes, it is more appropriate to use *AprioriHybrid* instead of *Apriori* Algorithm.

Usman *et al.* [4] proposed a model for multi-level mining of association rules using data warehouse schema. Clusters are created at multiple levels in the hierarchy and top-ranked dimensions are extracted to develop a schema for the data warehouse. The data warehouse schema is used to extract hidden patterns in the form of association rules which are then passed to the graphical component for visualization. The implementation of this model is not provided to show the effectiveness.

Hahsler & Karpienko [20] emphasize that the existing techniques for visualization of association rules are not able to handle situations where association rules are large in numbers. In this context, scatter plots, parallel plots and decker plots have been used by authors in the past; however, these do not handle large number of rules. In this research work, a new method is introduced called group matrix-based visualization. The technique uses clustering in order to create groups. The clusters are used for further exploration for individual rules. A plot format is introduced to show the interestingness of extracted association rules using colours and positioning of elements. However, the technique doesn't have option to explore rules using interestingness measures. The methodology has been tested using a small dataset and authors intend to evaluate this technique in future on a large dataset. Another issue with the methodology is the usage of R-Language which requires technical knowledge for exploration of association rules.

In a recent study done by Jeon *et al.* [21], social data is used for conducting a rule-based topic trend analysis. On line Analytical Processing (OLAP) has been used with Association Rule Mining (ARM) to perform the analysis. Twitter stream data has been used in order to test the methodology. Data is purified in the first step by removing un-necessary words. The second step involves the implementation of LDA which extracts topics from the data. Afterwards STAR scheme is generated and data warehouse is filled with the data from the dataset in this schema. Aggregate data within the data warehouse is used to extract patterns in the form of association rules. Although the patterns are extracted from a data warehouse, the rules are evaluated using conventional measures like support and confidence. Another issue with the technique is that, it creates only one fact variable, limiting the analysis in case where multiple fact variables should be created.

In this section, a literature review of previous techniques involving pattern extraction in large sets is discussed. We critically evaluate these techniques in the next section.

CRITICAL EVALUATION

This section provides the critical evaluation of the previous work done in the area of pattern extraction in large datasets. We present the summary of techniques in Table 1, by briefly discussing the major contributions in the particular study.

From the previous studies, it is observed that authors have emphasized on usage of dimension reduction methods, multi-level mining, evaluation with advanced measures and visualization support in mining over large datasets.

In recent techniques authors have used Agglomerative Hierarchical Clustering or K-Means clustering to create clusters at multiple levels before the mining process begins. Hierarchical clustering doesn't handle the data in large volume effectively [10]. Moreover, in this technique, if misclassification occurs at a stage, it cannot be corrected in further stages. Similarly K-Means clustering technique doesn't tackle the data in large size effectively. Another disadvantage of AHC is that it only works with numeric data. If the data contains mixture of variables, and variables are inter-related, such information is not used during clustering process. So in order to create cluster at different levels, techniques like SBAC can be used which use both nominal and numeric data while performing clustering.

Secondly, dimension reduction has been emphasized a lot in general for mining. In case of mining at multiple levels, dimension reduction has been done separately for numeric and nominal data. Due to this, semantic relationship between both types of variables is not kept while ranking process is applied. Thus ranking process needs to be applied at once instead of being applied separately for both types.

Thirdly, *Apriori* algorithm is not be a better choice than *AprioriHybrid* in case of large number of dimensions as suggested by authors previously. As per discussion in Agrawal *et al*. [14], *AprioriHybrid* suites in case of large datasets and combines *Apriori* and *AprioriTid* algorithms to create a hybrid approach for large datasets.

Fourthly, most of the authors have used conventional measures of interestingness like *Support* and *Confidence*. Some of the authors have used *Count, Min* and *Max* measures. A small number of authors have used diversity criteria for summarized data. To the best of our knowledge, Conciseness, Generality, Peculiarity and Unexpectedness have not been checked in multi-level mining in data warehouses. It is suggested that advanced measures of interestingness are used for evaluation.

Finally, there are some approaches which provide the visualization of association rules, but these techniques have problems with luminosity, colors and shape sizes in case of large number of resultant rules. It is suggested that an effective approach for visualization of rules should be developed in case where there is a large number of resultant rules.

Now we present a conceptual model which works in a multi-step framework, and handles the problems identified in previous literature. The implementation of this model is not covered in this study up-to date.

PROPOSED MODEL

In this section we present a model which aims at combining different techniques to create a framework for pattern extraction in large datasets. The process starts by taking data for mining process. It is assumed that dataset contains a mixture of nominal and numeric variables. We propose to create clusters of data at different levels of hierarchy using SBAC. Usage of SBAC ensures that both nominal and numeric variables are used for clustering. After the clusters are obtained, only top-ranked variables should be used for mining purposes in the second step. We propose to convert nominal variables to numeric variables and apply a single method to pick top ranked dimensions.

Principal Component Analysis (PCA) is proposed to be applied in this regard. After top ranked variables are obtained, a data warehouse schema is generated. The schema in this case consists of a fact table and multiple dimension tables. Data is shifted from dataset to data warehouse. We propose to apply *AprioriHybrid* to extract association rules for each cluster. Since the extracted rules are based upon aggregated data in the data warehouse, we propose to use advanced interestingness measures like Conciseness, Generality, Peculiarity and Unexpectedness to evaluate these rules. We propose to provide a visualization component for interactive analysis of extracted patterns.

TABLE I: Summary of techniques used for pattern extraction in large datasets

Title	Proposed Benefits	Dataset Used
Visualizing association rules in hierarchical groups. - [20]	Group-Matrix based visualization using clustering methods.	Yes
An improved Apriori algorithm for mining association rules - [15]	Improved version of Apriori in terms of time consumption.	Yes
Rule-Based Topic Trend Analysis by Using Data Mining Techniques. - [21]	Usage of OLAP with ARM for pattern extraction from twitter stream data.	Yes
A conceptual model for multi-level mining and visualization of association rules. -[4]	Presented a conceptual model that works at multiple levels in the hierarchy to mine association rules.	No
Diverse Association Rule mining through Statistical and Data mining Techniques. - [8]	Worked on ARM in data warehouse environment by creating STAR Schema. Dimension reduction is done using statistical measures. Advanced evaluation is done using Rae, CON and Hill measures.	Yes
Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. - [16]	Applied association rule mining for binary attributes and used advanced measures like Lift, OM and Lovinger etc along with provision of visualization support.	Yes

Mining association rules in very large clustered domains. - [12]	Provided Association Rule Mining concept at a group level.	Yes
Association Rule Mining on data cubes with Visualization ability. -[7]	Provided a framework OLEMAR to mine association rules. Visualization is provided using different shapes, colors and Semiology Principle.	Yes
Association Rule mining on data cubes - [2]	Association rule mining on data cubes using meta-rule guided approach. Evaluation is done using advanced measures (Lift, Loevinger)	Yes
Hierarchical grouping of association rules and its app. to a real-world domain -[18]	Worked on association rule mining at multiple hierarchy levels by creating groups at each level.	Yes
Discovering the association rules in OLAP data cube with daily downloads of folklore materials - [19]	Performed pattern extraction on specific data cubes called *FolkloreCubes*. Visualization support is provided using Column and Bar Graphs.	Yes
Survey of clustering algorithms - [11]	Provided comparisons of different clustering algorithms for high dimensional data	No
Unsupervised learning with mixed numeric and nominal data - [10]	Proposed Similarity-based Agglomerative Clustering (SBAC) – overcame the limitations of AHC	Yes
Association rule mining on data cubes and Visualization of Rules- [1]	Defined categories of associations in a multi-dimensional environment. A visualization component is provided using Column and Bar Graphs.	Yes
From data mining to knowledge discovery in databases. - [9]	Highlighted the importance of dimension reduction for mining process in large datasets	No
Sampling large databases for association rules - [17]	Presented a methodology to sample a large dataset to reduce the I/O size for mining process.	Yes
Fast Discovery of Association Rules. - [14]	Presented variant, *AprioriHybrid*, of previous algorithm for better performance.	Yes
Mining association rules between sets of items in large databases -[13]	Presented well-known association rule mining algorithm called *Apriori* for large datasets	Yes

The model (Figure 1) achieves its objective of mining at multiple levels using SBAC which creates clusters using both numeric and nominal data. Semantic relationship between variables is kept by applying a single method of ranking which is important to keep during ranking process. Moreover, patterns are proposed to be extracted using a better mining algorithm than previous studies. We propose to use *ApprioriHybrid*. The interestingness is evaluated using advanced measures since the data is aggregated during the mining process. The model also proposes to provide the ability to visualize patterns at the end of analysis process.

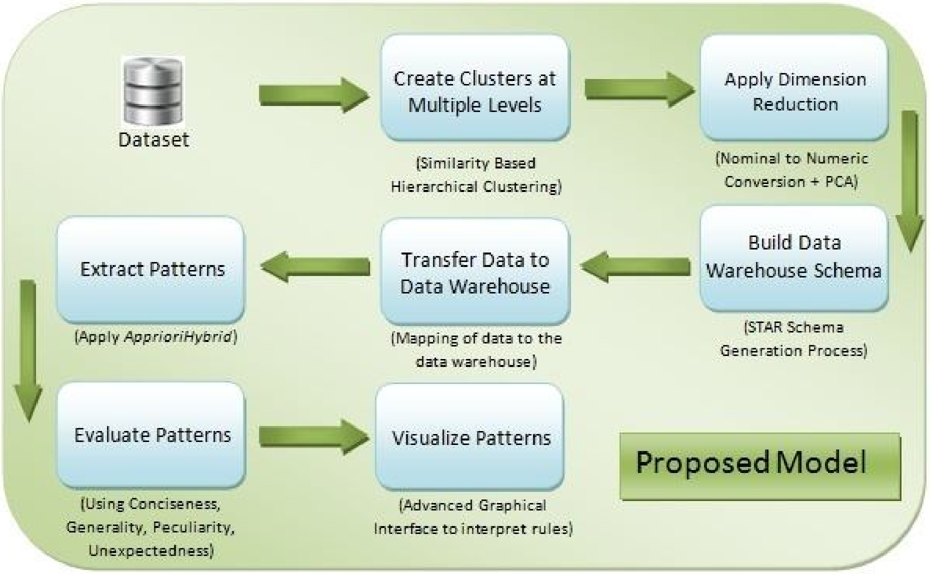
Dataset
Create Clusters at Multiple Levels
(Similarity Based Hierarchical Clustering)
Apply Dimension Reduction
(Nominal to Numeric Conversion + PCA)
Build Data Warehouse Schema
(STAR Schema Generation Process)
Transfer Data to Data Warehouse
(Mapping of data to the data warehouse)
Extract Patterns
(Apply *ApprioriHybrid*)
Evaluate Patterns
(Using Conciseness, Generality, Peculiarity, Unexpectedness)
Visualize Patterns
(Advanced Graphical Interface to interpret rules)
Proposed Model

Figure 1: Proposed Theoretical Model for Pattern Extraction in Large Datasets

CONCLUSION AND FUTURE WORK

In this research study, literature related to pattern extraction in large datasets has been reviewed. From the critical review, it has been found that there is a need to apply modern algorithms for multi-level mining, robust dimension reduction technique, robust pattern extraction algorithm, advanced evaluation measures and interactive visualization support for pattern extraction techniques in large datasets. In order to overcome the problems found, a theoretical model has been presented. The implementation of this model on datasets is suggested as a future work.

REFERENCES

[1] H. Zhu, “On-line analytical mining of association rules,” M.S. Thesis, School of Computing Science, Simon Fraser University, Canada. 1998 [Online]. Available: https://pdfs.semanticscholar.org/2e99/73e3805e8642b785e95b86d08e95370dae45.pdf

[2] R. B. Messaoud *et al*., “Enhanced mining of association rules from data cubes,” in *Proc. 9th ACM Int. workshop Data Warehousing OLAP*, 2006, pp. 11–18 [Online]. Available: https://www.researchgate.net/profile/Sabine_Loudcher/publication/220933876_Enhanced_Mining_of_Association_Rules_from_Data_Cubes/links/02e7e525bbc8265e7e000000/Enhanced-Mining-of-Association-Rules-from-Data-Cubes.pdf

[3] J. Han *et al.*, *Data mining: Concepts and Techniques*. Elsevier, 2011.

[4] M. Usman *et al*., "A conceptual model for multi-level mining and visualization of association rules," in *9th Int. Conf. Digital Inf. Manag. (ICDIM)*, 2014, pp. 175–181.

[5] A. Cuzzocrea, "Multidimensional mining of big social data for supporting advanced big data analytics," in *40th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO),* 2017, pp. 1337–1342 [Online]. Available: http://docs.mipro-proceedings.com/miprobis/bis_001_4618.pdf

[6] S. Ramirez-Gallego *et al.*, "A survey on data preprocessing for data stream mining: current status and future directions," *Neurocomputing*, vol. 239, pp. 39–57, 2017 [Online]. Available: https://www.researchgate.net/profile/Bartosz_Krawczyk/publication/313740355_A_survey_on_Data_Preprocessing_for_Data_Stream_Mining_Current_status_and_future_directions/links/59cb1ec60f7e9bbfdc36c013/A-survey-on-Data-Preprocessing-for-Data-Stream-Mining-Current-status-and-future-directions.pdf

[7] R. Ben Messaoud *et al*., "OLEMAR: an online environment for mining association rules in multidimensional data," in *Data mining and Knowledge discovery technologies*, IGI Global, 2007, pp. 1–35 [Online]. Available: https://hal.archives-ouvertes.fr/file/index/docid/476503/filename/adwm07_version_auteur.pdf

[8] M. Usman *et al*., "Discovering diverse association rules from multidimensional schema," *Expert Syst. Appl.*, vol. 40, no. 15, pp. 5975–5996, 2013 [Online]. Available: https://aut.researchgateway.ac.nz/bitstream/handle/10292/5460/ESWA.pdf?sequence=11&isAllowed=y

[9] U. Fayyad *et al.*, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, p. 37, 1996 [Online]. Available: https://vvvvw.aaai.org/ojs/index.php/aimagazine/article/download/1230/1131

[10] C. Li and G. Biswas, "Unsupervised learning with mixed numeric and nominal data," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 4, pp. 673–690, 2002.

[11] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, 2005 [Online]. Available: http://scholarsmine.mst.edu/cgi/viewcontent.cgi?article=1763&context=ele_comeng_facwork

[12] A. Nanopoulos *et al.*, "Mining association rules in very large clustered domains," *Inf. Syst.*, vol. 32, no. 5, pp. 649–669, 2007.

[13] R. Agrawal *et al.*, "Mining association rules between sets of items in large databases," *ACM Sigmod Record*, 1993, vol. 22, no. 2, pp. 207–216 [Online]. Available: http://www.it.uu.se/edu/course/homepage/infoutv/ht08/agrawal93mining.pdf

[14] R. Agrawal *et al.*, “Fast discovery of association rules.,” *Adv. Knowl. Discov. Data Min.*, vol. 12, no. 1, pp. 307–328, 1996 [Online]. Available: https://www.cs.helsinki.fi/u/htoivone/pubs/advances.pdf

[15] X. Yuan, “An improved Apriori algorithm for mining association rules,” in *Conf. Proc. AIP*, 2017, vol. 1820, no. 1, p. 80005 [Online]. Available: http://aip.scitation.org/doi/pdf/10.1063/1.4977361

[16] M. Plasse *et al.*, “Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set,” *Comput. Stat. Data Anal.*, vol. 52, no. 1, pp. 596–613, 2007 [Online]. Available: https://163.173.228.40/fichiers/RC1172.pdf

[17] H. Toivonen, “Sampling Large Databases for Association Rules,” *Proc. 22th Int. Conf. Very Large Data Bases*, vol. 96, pp. 134–145, 1996 [Online]. Available: http://www.cs.bilkent.edu.tr/~guvenir/courses/CS558/SeminarPapers/Sampling.pdf

[18] A. An *et al.*, “Hierarchical grouping of association rules and its application to a real-world domain,” *Int. J. Syst. Sci.*, vol. 37, no. 13, pp. 867–878, 2006.

[19] G. Bogdanova and T. Georgieva, “Discovering the association rules in OLAP data cube with daily downloads of folklore materials,” in *Int. Conf. Comput. Syst. Technol.*, 2005, vol. 6 [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.6435&rep=rep1&type=pdf

[20] M. Hahsler and R. Karpienko, “Visualizing association rules in hierarchical groups,” *J. Bus. Econ.*, vol. 87, no. 3, pp. 317–335, 2017 [Online]. Available: https://link.springer.com/content/pdf/10.1007%2Fs11573-016-0822-8.pdf

[21] Y. Jeon *et al.*, “Rule-Based Topic Trend Analysis by Using Data Mining Techniques,” in *Advanced Multimedia Ubiquitous Eng.*, Springer, 2017, pp. 466–473.