International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 6 665 – 670

Squential Step Towards Pattern Warehousing

Harshita Jain, M.tech Scholar, Madhav Institute of Technology & Science, Gwalior
Dr. Akhilesh Tiwari, Associate Professor, Madhav Institute of Technology & Science, Gwalior

Abstract: With the massive increase in the data, the demand by the analysts hyped for the proper repositories where they could analyse the concerned specific data patterns in order to make smart and quick decisions for the welfare and benefit of the business, organization or some social work. Pattern warehouse proved to be the best solution. This paper focuses on the discussion of existing architecture and moreover on the algorithms that is needed for retrieving the optimal patterns from the pattern warehouse. It also includes the detailed study about the sequential emergence of association rule algorithms which initially derive out patterns and later on those patterns are being optimized according to the interest of the analyst.

***Key Words**: Data Mining, Data Warehousing, Optimization, Patterns, Pattern Warehousing, Association Rule Mining.*

*****

I. INTRODUCTION :

Now-a-days, technological era of the world generates the immense amount of data gathered by systems that has needed to analyze as well as discover interesting information from such huge amount of data. Technological advancements and available storage perquisites are accountable for such explosive data. So there is urgent need for the development of tools and techniques regarding the analysis of such immense data. Data mining emerged as the new research area to meet this defiance.

Data mining (DM), also called **Knowledge mining, Knowledge extraction, Data/Pattern analysis, Data archaeology, Data dredging**, is the process for the extraction of valuable information from the huge amount of data. Data Mining is one of the most constitutive analysis step of the "Knowledge Discovery in Databases" or KDD Process. The term KDD is used for the explorative process of extraction of knowledge from data. An official definition of KDD given by Usama Fayyad in 1996 is: "KDD or Data Mining is non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data". The term data mining, a subset step of KDD Process, is used exclusively for the discovery as well as analysis stage of KDD process. Data mining is focused because it is most time consuming as well as most significant among KDD steps in the current scenario

As above discussion regarding immense data dictates that the popularity and importance of data mining has originated for two grounds: exponentially increasing volume of data and computational power. For example business activities produce an exponentiation stream of data (transactions) that is stored in larger and cheaper storage facilities. The consequence of the increasing stream of data and computational power create a motivation to develop data mining applications to extract novel, potentially useful as well as ultimately understandable knowledge from large volume of data.

Analyzing the huge amount of data is so tedious task in current scenario. For analyzing such immense data different data mining tasks are available like Classification, Clustering, Association rule mining (ARM), Prediction, and Regression etc [18]. Among these tasks Association rule mining is of our interest. Explorative discussion about association rule mining is done in the further section of this thesis but one of the most influential applications of association rule mining is Market Basket Analysis (Super Market Analysis) in the business perspective in which profit is the most spectacular measure in this economic era. Association rule mining technique extracts many rules from large transactional databases but only few rules are useful as well as profitable in the business point of view. So there is an urgent requirement to develop such framework for the generation of profitable as well as optimized rules. In different kinds of information databases, such as scientific data, medical data, financial data, and marketing transaction data; analysis and finding critical hidden information has been a focused area for researchers of data mining. How to effectively analyze and apply these data and find the critical hidden information from these databases, data mining technique has been the most widely discussed and frequently applied tool from recent decades [19]. Although the data mining has been successfully applied in the areas of scientific analysis, business application, and medical research and its computational efficiency and accuracy are also improving, still manual works are required to complete the process of extraction. Association rule mining model among data mining several models, including association

 ISSN: 2321-8169
665 – 670

rules, clustering and classification models, is the most widely applied method. The Apriori algorithm is the most representative algorithm for association rule mining [20]. It consists of many modified algorithms that focus on improving its efficiency and accuracy. However, two parameters, minimal support and confidence, are always determined by the decision-maker him/herself or through trial-and-error; and thus, the algorithm lacks both objectiveness and efficiency.

II. BACKGROUND

A. DATA WAREHOUSE

Data warehouses are used to consolidate data located in disparate databases. A data warehouse stores large quantities of data by specific categories so it can be more easily retrieved, interpreted, and sorted by users.

Warehouses enable executives and managers to work with vast stores of transactional or other data to respond faster to markets and make more informed business decisions. It has been predicted that every business will have a data warehouse within ten years. But merely storing data in a data warehouse does a company little good.

Companies will want to learn more about that data to improve knowledge of customers and markets. The company benefits when meaningful trends and patterns are extracted from the data.

B. ISSUES RELATED TO DATAWAREHOUSE:

- The size of single data warehouse was quite large. So it becomes tedious task to handle the management of data warehouse.
- For analysis purpose business analyst demands the consolidated information.
- Exponential increase in data day by day and the storing cost does not hold data warehouse as the best solution for the problem .
- Desired patterns are in volatile form in data warehouse, so even for small analysis the whole process of data mining has to be performed for obtaining certain results.

C. ADVENT OF PATTERN WAREHOUSE :

- As the size of the data warehouse is growing due to massive increase of data, business analyst are now not in the need of huge analytical data but they are interested in getting only the relevant patterns hidden within repositories.

D. PATTERN WAREHOUSE AND PATTERN MINING :

- Pattern warehouse is a kind of repository which stores the relevant patterns which are the representative of the relationship that exist between the data elements.
- Pattern mining is performed upon the patterns stored in pattern warehouse for generating analytical outcomes. Through pattern mining the analyst has to deal with small amount of information[7]

E. PATTERNS :

A set of items, subsequences or substructures that occur frequently together in a data set. It represents intrinsic and important properties of datasets.
It forms foundation for :

- Correlation
- Causuality analysis
- Mining sequential structure patterns
- Spatio temporal data
- Multimedia & data stream
- Classification which uses discriminative pattern based analysis
- Clustering uses pattern based subspace clustering

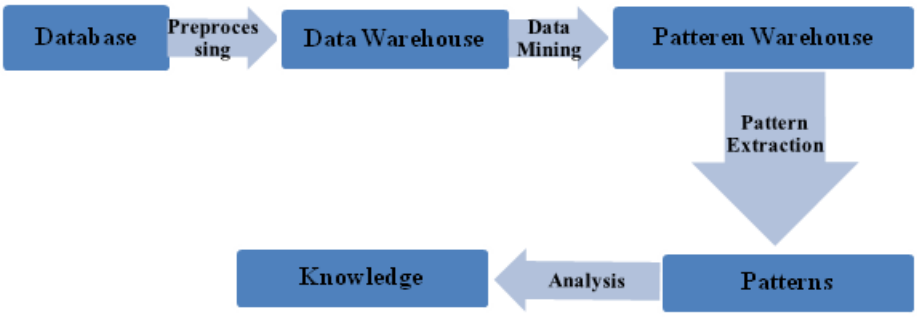

Figure 1: Flow Diagram for working of Pattern Warehouse

F. ASSOCIATION RULES:

Agrawal et al. first proposed the drawback of the miningassociation rule in 1993. They pointed out that some hiddenrelationships exist between purchased items in transactionaldatabases. Therefore, mining results can help decision-makersunderstand customers' purchasing behavior. An association rule isin the form of X→Y, where X and Y represent Itemset(I), or products,respectively and Itemset includes all possible items (i1, i2, . . .,in). The general transaction database (D= {T1, T2, . . ., Tk}) can represent the possibility that a customer will buy product Y after buying product X and X∩Y. However, the mining association rule mustaccord with two parameters at the same time:

1. **Minimal support**: Finding frequent itemsets with their supportsabove the minimal support threshold.

$$Support(X \rightarrow Y) = \frac{\text{no. of transactions which contain X\&Y}}{\text{total no. of transactions in the database}} - (1)$$

2. **Minimal confidence**: Using frequent itemsets found in Eq. (1) togenerate association rules that have confidence levels above theminimal confidence threshold.

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{no. of transactions which contain X\&Y}}{\text{no. of transactions which contain X}} - (2)$$

G. ASSOCIATION RULE MINING:

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It analyzes and present strong rules discovered in databases using different measures of interestingness. Agrawal et al. first proposed the issue of the mining association rule in 1993. They pointed out that some hidden relationships exist between purchased items in transactional databases. Therefore, mining results can help decision-makers understand customers' purchasing behavior. An association rule is in the form of X→Y, where X and Y represent Item set (I), or products, respectively and Item set includes all possible items{i1,i2, . . .,im}. The general transaction database (D= {T1, T2, . . ., Tk}) can represent the possibility that a customer will buy product Y after buying product X . Based on the concept of strong, rules, Agrawal et al., introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy burger. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics.

H. SEQUENCING THE ASSOCIATION RULE ALGORITHMS:

The Apriori algorithm repeatedly generates candidate itemsets and uses minimal support and confidence to filter these candidate itemsets to find high-frequency itemsets. Association rules can befigured out from the high-frequency itemsets. The process of finding high-frequency itemsets from candidate itemsets. After the Apriori algorithm has generated frequent itemsets, association rules can be generated. As long as the calculated confidence of a frequent itemset is larger than the predefined minimal confidence, its corresponding association rule can be accepted. Since the processing of the Apriori algorithm requires plenty of time, its computational efficiency is a very important issue. In order to improve the efficiency of Apriori, many researchers have proposed modified association rule-related algorithms. The algorithm logically divides the database into a number of non overlapping partitions, which can be held in the main memory. The partitions are considered one at a time and all large itemsets are generated for that partition. These large itemsets are further merged to create a set of all potential large itemsets. Then these itemsets are generated.

Park et al. in 1995 proposed the DHP algorithm. DHP can be derived from Apriori by introducing additional control. To this purpose, DHP makes use of an additional hash table that aims at limiting the generation of candidates as much as possible. DHP also progressively trims the database by discarding attributes in transactions or even by discarding entire transactions when they appear to be subsequently useless.

Toivonen proposed the sampling algorithm in 1996. This algorithmis involved in finding association rules to reduce databaseactivity. The sampling algorithm applies the level-wise methodto the sample, along with a lower minimal support threshold,to mine the superset of a large itemset.

The dynamic itemset counting (DIC) algorithm. The DIC algorithmwas proposed by Brin et al. in 1997. One of the main designmotivations was to limit the total number of passes performed overdatabases. DIC partitions a database into several blocks marked bystart points and repeatedly scans the database. In contrast to Apriori,DIC can add new candidate itemsets at any start point, instead ofjust at the beginning of a new database scan.

The Pincer-Search algorithm was proposed by Lin et al. in 1998 and can efficiently discover the maximum frequent set. The Pincer-Search algorithm combines both the bottom-up and top-down directions. It is used to prune candidates in the bottom-up search. Another very important characteristic of the algorithm is that it is not necessary to explicitly examine every frequent itemset. Therefore, it performs well even when some maximal frequent itemsets are long. The Pincer-Search algorithm can reduce both the number of times the database is readand the number of candidates considered.

An efficient hash-based method for discovering the maximal frequent set (HMFS) algorithm. In 2001, Yang et al. proposed the efficient hash-based method, HMFS, for discovering maximal frequent itemsets. The HMFS method combines the advantages of both the DHP and the Pincer-Search algorithm. The combination of the two methods leads to two advantages. First, the HMFS method, in general, can reduce the number of database scans. Second, the HMF Scan filter the infrequent candidate itemsets and use the filter edit itemsets to find the maximal frequent itemsets.

Genetic algorithms have also been applied in association rule mining. This study uses weighted items to represent the importance of individual items. These weighted items are applied tothe fitness function of heuristic genetic algorithms to estimate the value of different rules. These genetic algorithms can generate suitable threshold values for association rule mining. In addition, Saggar et al. (2004)[21] proposed an approach concentrating on optimizing the rules generated using genetic algorithms. The most important aspect of their approach is that it can predict the rule that contains negative attributes [10]. In another study, a genetic algorithm was employed to mine the association rule oriented to the dataset in a Manufacturing Information System (MIS). According to the testresults, the conclusion drawn stated that the genetic algorithm had considerably higher efficiency [11].

In another study, an ant colony system was also employed to data mining under multi-dimensional constraints. The computational results showed that the proposed method could provide more condensed rules than the Apriori method. In addition, the computation time was also reduced. In addition, this method was integrated with the clustering method to provide more precise rules. The improved algorithms described above have dramatically improved the efficiency of the Apriori algorithm. Among these improvements, some studies have focused on solving the problem of setting minimal support and minimal confidence to achieve more objective and efficient association rules. An increasing number of studies combine meta-heuristic methods, such as genetic algorithms and ant colony systems, with Apriori algorithm. These studies have proven that such integration can improve Apriori'sefficiency and discover association rules more precisely. However,the problems still exist.

III. RELATED WORK:

The recent approach by author Vishakha Agarwal (2016)[13] consist of an evolutionary algorithm (genetic algorithm) which works upon the optimization engine and generates optimal patterns from pattern warehouse.
The workflow to obtain optimal patterns is :

Pattern Warehouse → Optimization engine →Repository for Optimal Patterns

The work did not focussed on the computational efficiency and the automatic determination of threshold values.

Bartolini et al. (2003)[1], founded the architecture of pattern warehouse where the row data is collected from discrete sources and patterns obtained, which were non volatile and persistant, are stored dedicately. Meanwhile, this Pattern Base Management system was differentiated from DBMS in order to process patterns using query language. It consist of 3 layers, namely 1. Pattern layer which consist of pattern collection 2. Type layer clusters the patterns of similar type which can be built in or user defined 3. Class layer have the collection of semantically related patterns. Likewise, it provides analysis results to the end user which involves extension of SQL to retrieve patterns but ultimately did not hold as the sufficient implemention due to less emphasis on raw data behaviour and nature along with less tendancy to handle semantically rich patterns.

Bartolini et al, (2004)[3] developed a framework which compares patterns, that are grouped in "patterns" & "complex patterns", using similarity operator 'SIM'. SIM is formulized in a way that it uses simple patterns without considering complex patterns. It deals with how to reconcilethe structure and make them comparable. Meanwhile, it lacks the wide consideration of patterns to cover working of aggregation function with respect to combined structure and measure similarity. The work emphasis is less on operator's applicability for pattern retrieval.

Vassiliadis (2004)[2] introduced a generic schema that comes variety of patterns and operators detects the similarity among set of association rules with respect to decision tree. Operations used were similarity test, hypothesis testing, prediction of the future, cross over from patterns of data and classification was also suggested for data retrieval and patern processing. He has introduced only pattern representation issues and not any logical modelling that can be implemented.

Catania et al. (2004)[15] considered issues regarding variability of sources / raw data and also discussed several issues like heterogeneity, temporal, query language etc of pattern. It included the general Pattern Retrieval (PR) process to accommodate all kinds of patterns base on PANDA. Meanwhile it lacks the pattern validation in case when source data has been changed or updated. The work presented has little discussion on Temporal Pattern Manipulation Language (TPML) but lacks relation with PR.

Rizzi (2004)[7] provided foundation for design and implementation of pattern bases using UML (Unified Modelling Language). They addressed main issues in static modelling, representation of relationships b/w patterns and functional and dynamic modelling. It lacks the discussion on how patterns are distinguished according to static, dynamic and functional point of view. Authors introduced new pattern relationship such as specification composition & refinement. The operators needed to be introduced to carry out and find those relations. Raw data and source schema are paid less focus.

Kotsifakes et al. (2005)[14] introduced 3 well known database domain namely relational, object – relational and XML. These were compared based on criteria like generality, extensibility and querying effectiveness. Comparison shows semi structure (XML) is more appropriate for pattern base. The work is limited to pattern base. The work is limited to pattern representation only and needed to discuss PR process in detail and evolves indexing as an importance need for PR but lacks how indexing will work on patterns. Little emphasis on pattern storage schema. The work was also extended for query based retrieval method but it is limited to structure date and cannot fit on patterns efficiently.

Barbara Catania et al. (2005)[9] proposed a Pattern Management System for management of patterns. This architecture had 3 layers namely Physical layer, Middle layer and External layer. Pattern base is in lower most physical layer. Middle layer contains PBMS engine composed of all interpreters and query processors. External layer consist of the results extracted by making PML/PQL requests to the PBMS engine.

Terrovitis et al. (2006)[8] suggested conceptual design of pattern warehouse through ER - Model, Star Schema, Snowflake schema & Galaxy Schema. Pattern retrieval cannot be performed in similar manner as we perform query based.

Manolis et al. (2007)[4] discussedthe logical foundation and mapping that covers data, patterns and their intermediate mappings. The work does not cover important pattern retrieval part and allows designer to organise semantically similar patterns and can be subsequently queried.

Mazon et al (2008)[12] discussed that facts dimension hierarchy is important to explore the information at different levels of details. They represented a conceptual model to accommodate summarizability by adopting the normalization process.The work is more concentrated on normalization process rather than central issues of summarizability. The summarizability issue is important and its inadequate handling may cause to erroneous output of pattern aggregation.

Kotsifakos et al. (2008)[10] proposed Pattern – Miner. The architecture consist of many modules , a pattern base and DM engine. Modules execute specific tasks and pattern base acts like a repository of patterns. Modules are meta mining module which deals with pattern extraction, pattern monitoring module gives meta clustering results and pattern comparison module takes patterns as input and provides comparison results.

Tiwari et al. (2014)[11] presented an architecture named Pattern Warehousing Management System (PWMS) which consists of four major layers namely raw data layer, data mining engine layer, PBMS layer and application layer. Further it is divided into two tiers which are pattern tier that stores patterns and type tier that stores the patterns according to their respective types.

IV. CHALLENGING ISSUES :

With the launch of this new concept " Pattern warehouse", the consistent efforts are being taken to improve it in its architectural, structural and querying aspects. The efforts requires more to be on using algorithms which works upon the optimization engine and generates optimal patterns from pattern warehouse. Various challenges in order to strengthen the coined concept includes :

- The use of genetic algorithm misses the computational efficiency and the automatic threshold value set up is also the matter of concern. The requirements of parameter settings, like crossover and mutation, make the procedures more complicated.
- Minimal support and confidence, are always determined by the decision-maker him/herself or through trial-and-error; and thus, the existing algorithm lacks both objectiveness and efficiency.
- If the experience rule is employed during association rule threshold decisions, such as in the determination of minimal support and minimal confidence, experimental data stored in the test database are linear, as a result, these experiences cannot completely reflect the actual situation.
- More emphasized work is required to store data structure which can hold every type of patterns.

- Still the technique to filter out relevant and non relevant patterns is lacking.
- Attempt made to extend SQL to retrieve patterns is not sufficient especially semantically rich patterns.
- Updating the patterns and to make retrieval of patterns more feasible is still a topic which requires continuous betterment.
- Logical architecture of pattern warehouse still misses the repositories which are necessary in making the generalized architecture.

V. CONCLUSION:

The research work has shown that, though many proposals and work exist towards the improvement of pattern warehouse, but in terms of practical feasibility to represent structural repository and pattern retrieval is still missing. The work also discussed the need of proper optimizing algorithm for optimization engine which can generates optimal patterns from pattern warehouse. Meanwhile, it is justifiable to conclude that pattern warehouse is more easy for the users to use the patterns for their objective research

REFERENCES

[1]. Bartolini, I., Bertino, E., Catania, B., Ciaccia, P., Golfarelli, M., Patella, M., & Rizzi, S. (2003) "Patterns for next-generation database systems: PRELIMINARY results of the PANDA Project ". Proceedings of the Eleventh Italian Symposium on Advanced Database Systems, SEBD 2003, Cetraro (CS), Italy.

[2]. Vassiliadis, P.(2004) ' Panel : pattern management challenges', International Workshop on Pattern Representation and Management, PaRMa04, Heraklion – Crete, Greece.

[3]. Bartolini, I., Ciaccia, P., Ntousi,I., Patella, M. and Theodoridiss, Y. (2004) 'A unified and flexible framework for comparing simple and complex patterns', Proceedings of ECML – PKDD'04, LNAI 3202, Springer Berlin Heidelberg, pp. 496 -499.

[4]. Manolis, T., Vassiliadis, P. and Spiros, S.(2007) 'Modeling and language support for the management of pattern – bases', Data & Knowledge Engineering, Elsevier, Vol. 62, No. 2, pp. 368 – 397.

[5]. Tiwari Vivek, & Thakur, R.S.(2014). Pattern Warehouse : A dedicated pattern managemet system. Encyclopedia of Business Analytics and Optimization. Hershey, PA : Business Science Reference, pp. 1799 – 1808.

[6]. J. Han and M. Kamber, "Data mining: Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers, San Francisco, Elsevier, 2006.

[7]. S. Rizzi, "UML-based Conceptual Modeling of Pattern-Bases", In Proceedings of the International Workshop on Pattern Representation and Management, Heraklion, Hellas, 2004.

[8]. M. Terrovitis, P. Vassiliadis and S.Skiadopoulos, "Modeling and Language Support for the Management of Pattern-Bases", Data & Knowledge Engineering, Elsevier, pp: 368,397,2007.

[9]. B. Catania, A. Maddalena and M. Mazza, "PSYCHO : A Prototype System for Pattern Management ", Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005.

[10]. E. Evangelos and E. Kotsifakos, "Pattern-Miner: Integrated Management and Mining over Data Mining Models", KDD Proceedings of the 14th ACM SIGKDDinternational conference on Knowledge discovery and data mining, pp: 1081-1084, August 2008.

[11]. V. Tiwari and R. S. Thakur, " P2ms : A Phase – wise Pattern Management system for Pattern Warehouse" , In. J. Of Data Mining, Modeling amd Management, Inderscience, vol. 5, no. 3, pp : 1 – 10, July 2014.

[12]. Mazon J N, Lechtenborger J and Trujillo J 2008 Solving summarizability problems in fact- dimension relationships for multidimensional models. In : ACM 11th International Workshop on Data Warehousing and OLAP(DOLAP 08), Napa Valley, USA, PP : 57 – 64.

[13]. Vishakha Agarwal, Akhilesh Tiwari, " A Novel Optimal Pattern Mining Algorithm using Genetic Algorithm" In. J. Of Computer Applications(0975 – 8887), volume 144 – No. 4, June 2016.

[14]. Kotsifakos, E., & Ntoutsi, I.(2005). Database support for data mining patterns. In Proceedings of the 10th Panhellenic Conferences on Advances in Informatics (ACM – PCI' 05).

[15]. Catania b, Maddalena A, Maurizio M, Bertino E and Rizzi S 2004 A framework for data mining pattern management. In : Proceeding of 8th European Conference Knowledge discovery in database : PKDD, Pisa, Italy, 87-98, Springer, Berlin Heidelberg.

[16]. Turner RC, Holman RR, Matthews D, Hockaday TD, Peto J (1979). "Insulin deficiency and insulin resistance interaction in diabetes: estimation of their relative contribution by feedback analysis from basal plasma insulin and glucose concentrations.". Metabolism. 28 (11): 1086–96.

[17]. Badri Patel et. al. "Optimization of Association Rule Mining Apriori Algorithm Using ACO" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-1, March 2011

[18]. Pujari AK (2001) Data Mining Techniques. University Press .

[19]. Badhe V, Thakur RS and Thakur GS (2015) Vague Set theory for profit Patterns and decision making in uncertain data. International Journal of Advanced Computer Science and Applications, Vol. 6, No. 6.

[20]. Chen Y, Zhao Y and Yao Y (2007) A Profit-based Business Model for Evaluating Rule Interestingness. Proceedings of the 20th Canadian Conference on Artificial Intelligence (CAI'07). pp 296-307.

[21]. Manish Saggar, Ashish Kumar Agarwal and Abhimunya Lad, "Optimization of Association Rule Mining using Improved Genetic Algorithms"IEEE 2004.