A comprehensive review of pattern warehousing technique

Shaheen Usmani, Amit K. Manjhvar, Gaurav Sharma
Research scholar, Assistant Professor, Research Scholar
Madhav Institute of Technology & Science, Gwalior

Abstract - In today's world data is increasing rapidly in massive amount in every organization. Various kinds of new repositories are introduced as the demands of analysis of data are increased. This paper described the various repositories used for data storage and efficient management of data. And also expose the journey from file processing system to a new repository called as pattern warehouse.

Key Words - Data management, Data warehouse, Data Mining, Pattern Mining, Pattern Warehousing.

I. INTRODUCTION

now a days as data is growing very rapidly in every organization either it is private or government organization, and managing the huge amount of data is a challenging task. the repositories used for storage and management of data are as follows.

i. File Processing System

Before the use of computer a manual file system was used to maintain the records and files. All the data was stored in files but this system was good only for small organizations having small number of items. File processing systems was an early attempt to computerize the manual filling system. Basically a file system is a method for storing and organizing computer files and the data they contain to make it easy to find access them. File systems may use storage devices such as hard disk or CD-ROM and involve maintaining the physical location of the files.

Characteristics of File Processing System-

- It is a group of files storing data of an organization and each file is independent from another.
- Files are designed by using programs written in programming languages like as COBOL, C, C++.
- Each file contains data for specific area or department like library, student fees and examinations.

But traditional file system had many disadvantages or limitations

- Duplication of data
- Data dependence
- Difficulty in representing data from the user's view.
- Data inflexibility.
- Incompatible file formats
- Lack of data security
- Concurrency problem
- Poor data modeling of real data world
- Time consuming

To overcome these limitations of traditional file processing system database management system is used.

ii. Database Management System

Since the 1960's database and information technology has been evolving systematically from primitive file processing systems to sophisticated and powerful database systems. The research and development in database system since the 1970's has progressed from early hierarchical and network database systems to the development of relational database system.

Database is a collection of related data and data is a collection of facts and figures that can be processed to produce information. A database management system stores data in such way that it becomes easier to retrieve, manipulate and produce information. This database management system manages the database perfectly and also it manages the schema of the database and analyzes queries that will retrieve the efficient data from the database. Database systems are designed to manage large bodies of information, management of data involves both defining structures for storage of information and providing mechanism for the manipulation of information [1].

Characteristics of DBMS-

- Stores any kind of data
- Support ACID properties
- Represents complex relationship between data
- Concurrent use of data

- Defines structure of database
- Reduce redundancy
- Maintain consistency
- Provides security
- Reliable
- Stores data in the form of tables

But many organizations have their businesses at different places and needs to analyze data for business decision making.

To overcome this issue data needs to be gathered at one place instead of storing at discrete locations, which generated the need of a repository sufficient to store this data. This repository is called data warehouse.

iii. Data Warehouse

In 1970's firstly Bill Inmon defines the term data warehouse. The concept of data warehousing started in the late 1980's proposed by IBM researchers "Barry Devlin and Paul Murphy" they developed the "Business data warehouse".

The data warehousing concept was proposed to provide an architectural model for the flow of data from operational systems to decision support environments. Data warehouse integrate data from various discrete and autonomous data sources.

"A data warehouse is a subject oriented, integrated, time variant and collection of data" [2].

Data warehouse stores the huge amount of data and provides the multidimensional view of data to the business analyst.

[A]. Types of Data Warehouse

- **Enterprise Warehouse-** Covers all areas of interest for an organization.
- **Data mart-** Covers a subset of corporate wide data that is of interest for a specific user group.
- **Virtual Warehouse-** It offers a set of views constructed on demand on operational databases.

[B]. Data Mining

The process by which valuable information is being extract from the data warehouse is called data mining. Data mining plays a vital role in the current growing rate of data rapidly day by day.

Data mining is also called Knowledge mining, data/Pattern analysis, data dredging. Data mining is one of the most important steps in the process of knowledge discovery in databases.

According to Usama Fayyad "KDD or data mining is non-trivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data" [3]. The steps of KDD are as follows:

i. Data cleaning- In this step, the noise and inconsistent data is removed.
ii. Data integration- In this step multiple data sources are combined.
iii. Data Selection- In this step the data relevant to the analysis task are retrieved from the database.
iv. Data Transformation- In this step data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
v. Data Mining- It is the approach of extracting valuable information from huge amount of data. In this step intelligent method like as decision trees, classification, clustering are applied in order to extract patterns.
vi. Pattern Evaluation- In this step data patterns are evaluated.
vii. Knowledge representation- In this step knowledge is represented in an efficient manner.

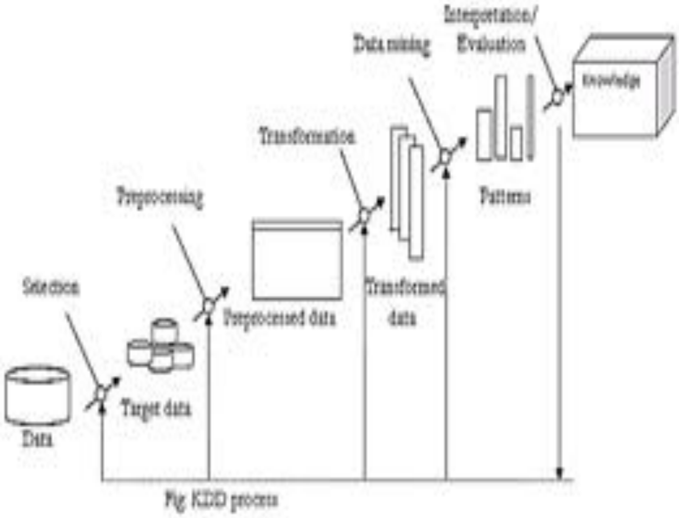

Figure 1: Knowledge discovery in database

[C]. Issues Related to Data warehouse

- The size of single data warehouse is very large so that the management of data warehouse is a difficult task.
- As the size of data warehouse is huge, it means it contains lots of data, but for analysis purpose analyst demands the consolidated and only useful 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 results. And it is very time consuming process.

Hence in order to deal with these issues of data warehouse a new repository is introduce and called as Pattern Warehouse.

iv. Pattern Warehouse

As the size of the data warehouse is growing due to massive increase in data day by day, business analysts are not demand the huge analytical data but they are interested in getting relevant and efficient patterns hidden within the repositories.

Pattern warehouse is a kind of repository which stores the relevant and useful patterns in an efficient manner.

A pattern is a set of items, subsequences substructures occur frequently in a data set. A pattern may be in the form of association rules, a decision tree, and a cluster of items [13].

e.g. {Milk}-> {Bread, Butter}

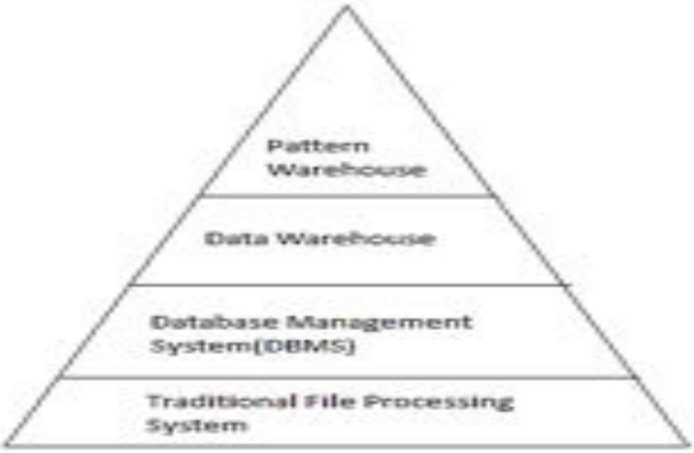
Pattern Warehouse
Data Warehouse
Database Management System(DBMS)
Traditional File Processing System

Fig2: Generation of Data storage Repository

In this figure the different kinds of repositories used for data storage from past decade to till now are shown.

II. Comparison among Database, Data Warehouse and Pattern Warehouse

Table 1.

Feature	**Database**	**Data Warehouse**	**Pattern Warehouse**
Characteristics	Operational Processing	Information processing	Analytical processing
Orientation	Transaction/application oriented	Analysis/Subject oriented	Analysis/subject oriented
User	Clerk, DBA, database professionals	Knowledge worker	Only Administrator
No. of user	Thousands	Hundreds	Few
Data	Current	Historical	Relational in the form of patterns
Summarization	Primitive, highly detailed	Consolidated	Consolidated
View	Flat relational	Multidimensional	Relational
Database design	E-R based	Star/snowflake	Still not present
Unit of work	Short, simple transaction	Complex query	Patterns
No. of records accessed	Tens	Millions	Hundreds
Size of repository	100 MB to GB	100 GB to TB	Few GB
Access	Read/Write	Mostly read	Read
Operations	Index/hash on primary key	Lots of scan	Less number of scan than data warehouse
Updates	Continuous	Periodical	Periodical
Query Language	SQL	DMQL	PMQL

III. Literature Review

In 2003 Bartolini et al. [4], proposed the architecture of pattern warehouse, and claimed the first time that the concept of pattern is a good candidate for generic representation. They preliminarily focused on pattern representation. In this raw data is collected from different sources and patterns obtained. They introduced a new idea of non-volatile or persistent pattern. This architecture mainly consists of three layers are as follows:

i. Pattern Layer- This layer consists of collection of patterns.
ii. Type Layer- This layer consists of cluster of patterns which are similar in type.
iii. Class Layer- This layer is the collection of semantically related patterns.

Similarly it provides analysis results to the end user which involves extension of SQL to retrieve patterns. But it did not hold sufficient implementation due to less emphasis on raw data behavior and with fewer tendencies to handle semantically rich patterns.

In 2004, Rizzi et al. [5], used UML (Unified modeling language) to model the pattern bases. They focused on the conceptual design of pattern bases by discussing how UML could be used. In this paper the pattern type, pattern class and specialization, composition and refinement also defined. It defines that modeling should be carried out along three coordinates- static, dynamic, and functional.

i. Static Modeling- Static modeling aimed at the describing the pattern base from a structural point of view.
ii. Functional Modeling- It is used to represent the processes used to establish relationship between data and patterns. In this type of modeling data flow diagram, use case diagram, flow charts or UML activity diagrams can be used.
iii. Dynamic Modeling- This aspect describes the how data and patterns are kept in synchronization. In this modeling state charts
used that represent the state of the pattern base in response to events and UML sequence diagram that is used to describe the sequence of message exchanged.

But some operators are needed to carry out and find relations among patterns. The pattern bases can also be represented without the use of usual syntax and semantics of UML.

In 2005, Evangelos & Irene [6] have studied the problem of the efficient representation and storage of patterns in a so called pattern base management system. The concept of pattern is the foundation of the pattern base management system. Evangelos defined three well known models from the database domain, the relational, the object relational, and the semi structured model. And these three models are compared based on the criteria and requirements like generality, extensibility, query effectiveness, pattern validation and pattern characteristics exploitation.

According to this comparison for the pattern representation problem a semi structured model is more appropriate than a relational or object relational schema.

But pattern base requires a data oriented approach and these approaches are structure oriented. Evangelos defined models only for pattern representation rather than to discuss any techniques for pattern retrieval.

In 2006, Terrovitis et al. [7], suggested the conceptual design of pattern warehouse for which they suggested models like E-R model, star schema, snowflake schema, and galaxy schema. They pointed out that pattern retrieval cannot be performed in a same manner as we perform query based extraction in data warehouse.

In 2008, Kotsifakos et al. [8], proposed the architecture of a tool namely called as pattern miner. The architecture consists of several modules a pattern base and data mining engine. Each module is responsible for executing specific and particular task and the resultant patterns are stored in the pattern base. Basically pattern base is a repository, various modules are Meta mining module, and pattern comparison module takes patterns as input and provides comparison results. The data mining engine helps the pattern miner to extract patterns. The analyst interacts with pattern miner engine to find out specific result.

In 2014, Vivek Tiwari et al. [9], designed a contextual snowflake modeling for pattern warehouse logical design by defining a context and kind of knowledge. Since the patterns are semantically rich and diverse therefore satisfying the user's interest is dependent on how and what kinds of pattern are being stored in a pattern warehouse. They described the four contexts of data.

i. Global data context- In this context patterns are created and stored in a pattern warehouse without concerning the domain of underlying raw data.
ii. Domain data context- In this context patterns are created and stored in pattern warehouse with concern domain of raw data.
iii. Scenario context- Patterns are created and stored in pattern warehouse with concern domain of underlying raw data and its scenario.
iv. Techniques and kind of knowledge context- Patterns are created and stored in pattern warehouse according to pattern retrieval techniques like association, clustering and classification patterns.

The author also described the four quality forms and also discussed and used the hybrid snowflake schema for pattern warehouse representation because inherently snowflake schema allows normalization.

In 2016, Vishakha Agarwal et al. [10], proposed a progressive step from a data warehouse to pattern warehouse. And conclude that the pattern warehouse is capable to store various patterns which are refined form of data and described that analysis task performed on pattern warehouse required less storage space as compared to data warehouse.

In 2016, Vishakha Agarwal et.al. [11], proposed a novel optimal pattern mining algorithm using genetic algorithm for finding the relevant and efficient patterns. In the proposed approach author is proposing an algorithm which works upon the optimization engine for generating optimal patterns from pattern warehouse. The steps of OPM-GA are

i. Select a pattern set
ii. Enter the value of crossover probability, mutation probability, and fitness value
iii. Then apply selection operator for selection of pattern.
iv. Perform crossover
v. Perform bitwise mutation
vi. Test the new offspring by calculating the fitness value
vii. If the pattern is efficient, place it into the optimal pattern set
viii. Else discard the pattern
ix. Repeat this process until whole pattern set is scanned.

In 2017, V. Tiwari et al. [12], described the important issues of pattern retrieval in pattern warehouse, and proposed an elementary view of pattern retrieval through classification. Pattern classification is not a single or independent process. It consists of mainly four parts:

i. Data to pattern conversion
ii. Make patterns persistent through pattern warehouse.
iii. Classify pattern in pattern warehouse
iv. Query processing on classified pattern for knowledge

But patterns are created through various data mining techniques, so it is very hard to develop pattern classification systems to classify patterns. The classification system should be able to identify the class boundary of attributes by only giving items and frequency count.

IV. Observations

There are various observations drawn according to literature review.

- As data warehouse have a multidimensional view the pattern warehouse does not have any view.
- Still there is no specific data structure define that can store all types of patterns.
- The patterns may be relevant and non-relevant. So there should be some technique to filter out non-relevant patterns.
- As data warehouse has schemas like star, snowflake and fact constellation, there should be schema in pattern warehouse.
- The use of genetic algorithm for pattern retrieval misses the computational efficiency and the automatic threshold value setup is also the matter of concern and choosing the crossover and mutation probability that will generate efficient pattern is also challenging task.
- The scope and objective of pattern warehouse must be clear.
- There is a requirement of technique for retrieval of efficient and optimal patterns.
- The architecture of pattern warehouse given by various authors but it still misses several necessary components and repository.
- Updating of pattern warehouse is also a task of concern.

V. Conclusion

This paper described the different kinds of repository used for storage and management of data. A pattern warehouse is a good repository in today's world according to analysis demand of analyst but still it is not free from some problems. According to this paper conclusion is that the pattern warehouse stores the patterns in an efficient and persistent manner. This paper also gives various observations about pattern warehouse and efficient pattern retrieval technique is required to retrieve the patterns.

References

[1] A. Silberschatz, H. F. Korth and S. Sudarshan, "Database System Concepts", Sixth Edition, McGraw-Hill, 2010.

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

[3] A. K. Pujari "Data Mining Techniques", Third edition, Universities press, India, 2007.

[4] Bartolini, I., Bertino, E., Catania, B., Ciaccia, P., Golfarelli, M., Patella, M. and Rizzi, S. 2003 Patterns for Next-generation Database Systems: Preliminary Results of the PANDA Project. In Proceedings of the Eleventh Italian Symposium on Advanced Database Systems, SEBD, Cetraro (CS), Italy.

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

[6] E. Kotsifakos and I. Ntoutsi, "Database Support for Data Mining Patterns", In Proceedings of the 10th Panhellenic Conference on Advances in Informatics, 2005.

[7] M. Terrovitis, P. Vassiliadis and S. Skiadopoulos, "Modeling and Language Support for the Management of Pattern Bases", Data & Knowledge Engineering, Elsevier, 2006.

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

[9] Tiwari and R. S. Thakur, "Contextual Snowflake Modeling for Pattern Warehouse Logical Design", Sadhana – Academy Proceedings in Engineering Science, Vol. 39, Springer, 2014.

[10] Vishakha Agarwal, Akhilesh Tiwari, "From Data Warehouse to Pattern Warehouse: A Progressive Step" International Journal of Engineering Research, Volume No.5, Issue No.4, pp: 249-252, April 2016.

[11] 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.

[12] Vivek Tiwari and Ramjeevan Singh Thakur, "Towards important issues of pattern retrieval: pattern warehouse" Int. J. Data Science, Vol. 2, No. 1, 2017.

[13] Harshita Jain, Akhilesh Tiwari, "Squential Step towards Pattern Warehousing" International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 6, ISSN: 2321-8169.

[14] Nidhi Tomar and Amit Kumar manjhvar: A survey on Data Mining optimization Techniques International Journal of Science Technology & Engineering, Dec -2015, Vol.2, Issue: 6, pp. 130-33, ISSN (online):2349-784X.

[15] Prateeksha Tomar, Amit Kumar Manjhvar, " survey report on various decision tree classification algorithm using weka tool", in International Journal of computer science and engineering", volume -5, Issue 3, March 2107.