Volume 4, Issue 1, January 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Changing Pattern of Warehousing: Data Warehouse to Pattern Warehouse

Dr. Ajay Kukshal
Computer Programmer
State Planning Commission
Government of Uttarakhand
Dehradun , India

Susheel Chandra Bhatt
Research Scholar,
Department of Computer Science
Kumaun University, Nainital,
Uttarakhand, India

Abstract -The field of database is an essential part of computing. Very large, highly structured repositories are increasingly replacing smaller, flat information spaces. Such repositories are often filled with multimedia documents, including image, audio and video data. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information from the rapidly growing volumes of digital data. The fact is that business users do not want large and heavy volume of data, but are interested in the patterns. This article represents a step toward a common framework that will ultimately provide a new vision and potential of Pattern Warehousing and discuss the conceptual framework of the Data Warehouse, Data mining , Pattern Warehouse and their utility.

Keywords: Data Warehouse, OLAP, Data Mining, Pattern Warehouse

I. INTRODUCTION

A database is a stored collection of data designed for efficient information retrieval, efficient data maintenance and centralized control of an organization's data. Data warehousing (DW) [1] is one of the fastest growing segments of the database management market, which, in turn, is one of the biggest markets for software. The area of data warehousing has grown out of a need for decision support that was not met by existing database management technologies. DW can be defined as 'subject-oriented', integrated, time-varying, non-volatile collections of data, that is used primarily in organizational decision making. Nowadays, it became an important strategy to integrate heterogeneous information sources in organizations, and to enable On-Line Analytic Processing (OLAP) [2]. A data warehouse is a collection of data, whereas an OLAP transforms data in a DW into strategic information. It can be normalized or renormalized. It can be a relational database, multidimensional database, flat file,hierarchical database, object database, etc. Data warehouses are run on client/server networks, which are mainly used for strategic applications [3].

Data warehousing and OLAP are essential elements of decision support which has increasingly become a focus of the database industry [4]. The OLAP is data summarization and aggregation tools that help data analysis. OLAP and Data Mining are disjoint, rather data mining is a step ahead from that of OLAP. Data mining allows automated discovery of implicit pattern and knowledge recognition in large amount of data [5]. Data Mining, also known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. This can be considered as the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses or other information repositories. [6]. Various application domains related with data management (storage, process, retrieval, data analysis) result in different forms of patterns representing the data insights.[7]

As a simple analogy, consider data as grapes and patterns of knowledge as wine -- data mining is then like the wine making process. While a data repository is a storage facility for grapes, a pattern repository is like a wine cellar. Data mining tools are then like wine making equipment. [8].

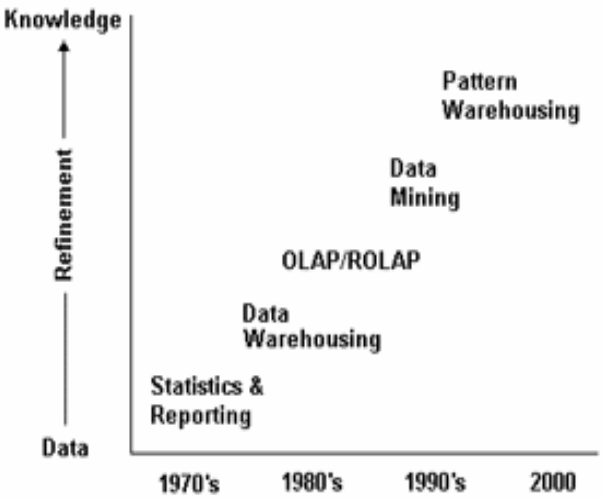
II. DATA WAREHOUSE TO PATTERN WAREHOUSE

Data warehousing ideas were recognized as early as the 1970s, but it wasn't until the mid 1980s that Bill Inmon coined the term data warehousing [9]. During the 1980s, businesses started using relational databases, which are primarily operational applications. Data warehouses do not serve only as information buffers for answering complex questions quickly but also as intermediate stages in the processing of information within the information system of an enterprise, where the information becomes more accurate and useful. It is a collection of technologies aimed at enabling the knowledge worker (executive, manager, analyst, etc) to make better and faster decisions.[2] It is just a frame, strategy, or solution to construct an integrated information infrastructure based on the traditional database management system. It provide a solid platform of integrated, historical data from which informational, analytical processing and analysis over a long historical time perspective can be done. It is the process to cleanse, filter, transform, summarize, and aggregate data generated in the operational environment and then put the data in a specially structured data warehouse for easy access and analysis by a disparate group of users [10].

A Data warehouse is a repository of integrated information available for querying and analysis [11]. It is an in-advance approach to the integration of data from multiple, possibly very large, distributed, heterogeneous databases and other information sources [12]. It can be seen as a set of materialized views defined over the sources. When a query is posed, it is evaluated locally, using the materialized views, without accessing the original information sources. The information stored at the Data warehouse can be used by organizations for decision support. [13].Even though there has been a lot of work on various aspects of materialized views with respect to Data warehouses, there is little or no theoretical work at all on providing a method for configuring a Data warehouse. As a consequence the design of a Data warehouse is haphazard and the quality of data is often dubious [14].

Data Mining and Knowledge Discovery in Databases (KDD) are rapidly evolving areas of research that are at the intersection of several disciplines, including statistics, databases, pattern recognition/AI, optimization, visualization, and high-performance and parallel computing [15]. It is concerned with the secondary analysis of large databases in order to find previously unsuspected relationships which are of interest or value to the database owners [16].

Data Mining is mainly concerned with methodologies for extracting data patterns from large data repositories. The extracted patterns are evaluated based on some interestingness measures that identify patterns representing knowledge. [7]. It holds dynamic information (which is automatically re-generated once a month with new data).



Progress of Knowledge Discovery [18]

In general, building data mining models is very expensive because it involves analyzing large amounts of data. It is universally acknowledged that the data pre-processing phase is the most consuming step in terms of computational time [17]. Pre-processing the data involves collecting, cleaning, transforming, preparing the raw data.

While SQL relies on the relational algebra, PQL uses the "pattern algebra". PQL was designed to access Pattern Warehouses just as SQL was designed to access databases. PQL allows knowledge-based queries just as SQL allows data based queries. PQL uses SQL as part of its operation, i.e. PQL queries are decomposed into a set of related SQL queries, then the results are re-combined. [18]. The most relevant research effort in the literature, concerning pattern management is found in the field of inductive databases, meant as databases that, in addition to data, also contain patterns [19] [20].

Information overloading is today a serious concern that may hinder the potential of modern web-based information systems [21].As a recent survey states [22] the world produces between 1 and 2 exabytes of unique information per year, 90% of which is digital and with a 50% annual growth rate. Clearly, this sheer volume of collected data in digital form calls for novel information extraction, management and querying techniques, thus posing the need for novel Database Management Systems (DBMSs).

A promising approach to deal with this problem is represented by knowledge extraction methods able to produce artifacts (also called patterns) that concisely represent data. Patterns are usually quite heterogeneous and voluminous. So far, little emphasis has been posed on developing an overall integrated environment for uniformly representing and querying different types of patterns.

Since patterns represent relevant knowledge, often very large in size, it is important that such knowledge is handled as first-class citizens. This means that patterns should be modeled, stored, processed, and queried, in a fashion analogous to data in traditional DBMSs [21].

III. WHY DO WE NEED PATTERN WAREHOUSE?

However, because knowledge is so much more compact than data, the Pattern Warehouse is only a fraction of the size of the data warehouse, allowing the patterns many years to be stored with ease, even when the data is no longer available [8].

In case of data analysis, when a user wants knowledge, analysis is performed. In case of knowledge access, analysis is automatically done beforehand, i.e. the “knowledge on demand”.

A successful pattern data base management system has great potential in applications in the area of Business, Trade, Services and Industry as we can see that in these areas we have a tremendous increase of data and very little offer from focused commercial products. The core algorithms are now a small part of the overall application and users do care about compact and rigid information which a Pattern-base Management System (PBMS) could offer. Also the results of the data mining process or of a dbase query will drive efforts in areas such as marketing, risk management, and credit scoring. Each of these areas is influenced by financial considerations that need to be incorporated in the focused data analysis process. A business user is concerned with maximizing profit, not minimizing a function error and thus the necessary information to make these financial decisions (costs, expected revenue, etc.) is often available and ideally should be incorporated in the pattern [23].

IV. CONCLUDING REMARKS

A Data warehouse is a repository of integrated information while Data Mining is mainly concerned with methodologies for extracting data patterns from data repositories and stored in Pattern Warehouse which is only a fraction of the size of the data warehouse. Stored pattern are use by business users such as marketing analysts, bank branch managers, store managers, etc.

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to Prof. Durgesh Pant for their valuable support to write this paper.

REFERENCES

1. W. K. Inmon. *Building The Data Warehouse*. John Wiley & Sons, 2nd ed. (1996).
2. Vassiliadis Panos, Data Warehouse Modeling and Quality Issues. Ph.D.Thesis.2000,
3. Fisher, Lawrence, .Along the Infobahn., *Strategy & Business*, Third Quarter,1996, http://www.strategybusiness.com/tecnology/96308/page1.html.
4. Chaudhuri Surjit, Dayal Umeshwar. *An Overview of Data Warehousing and OLAP Technology*, Appears in ACM Sigmod Record, March 1997, http://www.compapp.dcu.ie/~mark/TEACHING/CA-306/Chaudhuri.pdf
5. Data Mining, *Data Warehouse*, BPB Editorial Board ,BPB Publication, Pressworks New Delhi, First Edition, 2004,pp. 37
6. Osmar R. Zaïane, *Introduction to Data Mining*, CMPUT690 Principles of Knowledge Discovery in Databases, 1999
7. M.Vazirgiannis, M.Halkidi, D.A. Keim , I.Ntoutsi, A. Pilrakis, S. Theodoridis, Y. Theodoridis , G. Tsatsarois, E. Vrachnos . "*Recent Advances on Pattern Representation and Management*" PANDA Technical Report Series , 2003 , pp. 4– 5, http://dke.cti.gr/panda/
8. Park, Kim Youngwok. Cha Youngmin, Kim Seongbum ; *The Concept of Pattern Warehouse and Contemplate an Application in Integrated Network data Ware*, http://www.knom.or.kr/knomreview/v4n2/1.pdf
9. Watson, Hugh J, and Barbara J Haley.,Managerial Considerations., *Communications of the ACM*, 41(9), 32-37, (September 1998).
10. Shi Dongyuan, Lee Yinhong, Duan Xianzhong, Duan Q.H. , *Power System Data Warehouses*, IEEE, 2001,pp 50
11. J.Widom, editor. Data Engineering, Special Issue on *Materialized Views and Data Warehousing*, volume 18(2). IEEE, 1995.
12. J.Widom. *Research problems in data warehousing*.In Proc. CIKM, pages 25-30, Nov. 1995.
13. Theodoratos Dimitri, Sellis Timos. *Data Warehouse Configuration*. Proceedings of the 23rd VLDB Conference Athens, Greece, 1997,pp 126
14. Theodoratos Dimitri and Sellis Timos , *Proceedings of the 23rd VLDB Conference Athens, Greece, 1997*,pp 126
15. Bradley P. S., Fayyad Usama M., Mangasarian O. L., *Mathematical Programming for Data Mining: Formulations and Challenges*. Journal on Computing 11, 1999, 217-238
16. HAND David J., *Data Mining: Statistics and More?*. The American Statistician, May 1998 Vol. 52, No. 2
17. Cios, K., Pedrycz, W., Swiniarski, R. & Kurgan, L. (2007). *Data Mining: A Knowledge Discovery Approach*. Springer. 17, 22
18. *The data Warehouse vs The Pattern Warehouse*, 2003 ; An Information Discovery, Inc. White Paper. File://\\Server\warehousing\The Data Warehouse vs The Pattern Warehouse.htm
19. T. Imielinski and H. Mannila, "*A database perspective on knowledge discovery*," *Communications of the ACM*, vol. 39(11), pp. 58–64, 1996.
20. L. De Raedt, "*A perspective on inductive databases*," *SIGKDD Explorations*, vol.4(2), pp. 69–77, 2002
21. Terrovitis Manolis, Vassiliadis Panson, Skiadopoulos, Bertino Elisa, Catania Barbara, Maddalena Anna, Rizzi Stefano. *Modeling and language support for the management of pattern-bases*; ScienceDirect; Data & Knowledge Engineering 62(2007) 368-397
22. P. Lyman and H. R. Varian, "*How much information*," http://www.sims.berkeley.edu/how-much-info, 2000.
23. Martin Nelke (MIT). "*Recent Overview of Pattern Related Applications*" Document Information Record, 2008, pp. 3 , http://dke.cti.gr/panda/index.htm.

