ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 8, Issue, 05, pp.30470-30474, May, 2016

INTERNATIONAL JOURNAL
OF CURRENT RESEARCH

RESEARCH ARTICLE

DATA MINING FOR BUSINESS INTELLIGENCE WITH DATA INTEGRATION

[1,*]Mubeena Shaik, [2]Dr. Wali Ullah, [3]Dr. Sheela Rani, C. M.

[1]Research Scholar, Department of Computer Science, KL University, India
[2]Assistant Processor, Department of Computer Science, Jazan University, KSA
[3]Associate Professor, Department of Computer Science, KL University, India

ARTICLE INFO

Article History:
Received 15th February, 2016
Received in revised form
27th March, 2016
Accepted 16th April, 2016
Published online 10th May, 2016

Key words:

Data Mining,
Data Integrity,
Knowledge Discovery,
Business Intelligence,
Dynamic Market.

ABSTRACT

Business Intelligence (BI) allows an Organization's executives to obtain a better understanding of their customers, sales, market, supply and resources, and competitors in order to make effective strategic decisions. The rapid growth of computer in today's business environment has increased the demand and urgency of successful business organization to be able to react quickly to the changing dynamic market demands in locally as well as globally.

Citation: Mubeena Shaik, Dr. Wali Ullah, Dr. Sheela Rani, C.M. 2016. "Data mining for business intelligence with data integration", *International Journal of Current Research*, 8, (05), 30470-30474.

INTRODUCTION

In general, data integration of multiple information systems aims at combining selected systems so that they form a unified form a new system and give users the vision of interacting with one single information system. The reason for integration is for two reasons. The first is given a set of existing information systems, an integrated view can be created to facilitate information access and reuse through a single information access point. And the second is given a certain information need, data from different complementing information systems is to be combined to gain a more comprehensive basis to satisfy the need. In the view of business intelligence context, the integration problem is commonly referred to as enterprise integration (EI). Enterprise integration denotes the capability to integrate information and functionalities from a variety of information systems in an enterprise. In this paper we are focusing on the challenges of data integration and finding the solution with the mining models for Enterprise Integration.

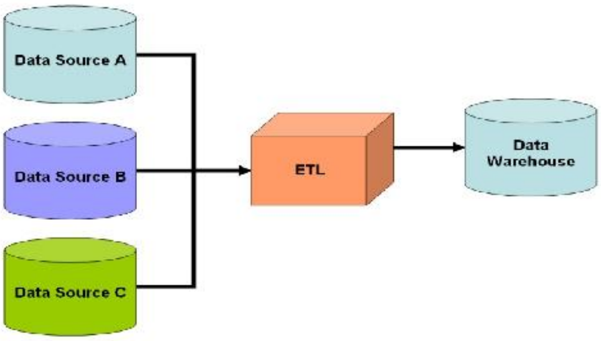
Data Mining in Business Intelligence

Data mining is a process of discovering meaningful correlations, patterns, constraints and rules by sifting through large amounts of data stored in warehouses. Data mining defines the pattern recognition techniques, statistical and mathematical formulae to recognize the suitable information (Surajit Chaudhuri, 1998). BI is a broad field and it is viewed differently by different people. Organizations are being capture, understand, their data to support decision making in order to improve business operations.

Data Integration

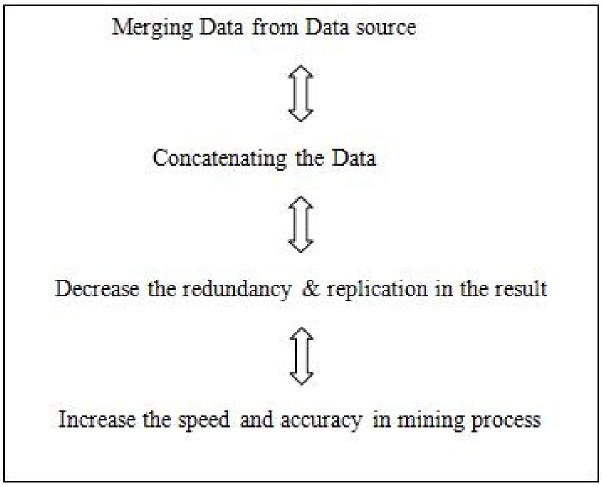
Data Integration is the process of integrating data from multiple sources and probably has a single view over all these sources. And answering queries using the combined information. Data integration becomes gradually important in cases of merging systems of two Organizations applications within one to provide a unified view of the organization's data assets, which is termed as data warehouse. Probably the most well known implementation of data integration is building an enterprise's data warehouse.

***Corresponding author: Mubeena Shaik,**
Research Scholar, Department of Computer Science, KL University, India.



The data warehouse enables a business to perform analyses based on the data in the data warehouse. This would not be possible to do on the data available only in the source system (John Clear, 1999). The reason is that the source systems may not contain corresponding data, even though the data are identically named, they may refer to different entities. ETL (Extract-Transform Load process) is totally performed outside of warehouse. Whereas warehouse only stores data. Data integration may involve inconsistent data and therefore needs data cleaning

The process of data integration as follows


Merging Data from Data source
Concatenating the Data
Decrease the redundancy & replication in the result
Increase the speed and accuracy in mining process

Data Integration Methods

Structural Integration: The purpose of structural integration is to try and resolve a variety of conflicts with regards to the structure of the schema. The schema of two data sources may suffer from structural heterogeneity for a number of reasons. Problems that may need resolution include that of type conflicts; for example an address attribute in one schema may have been represented as a string, whereas, in the second schema it has been defined as a struct. Further structural issues that may arise are naming inconsistencies, where attributes that define the same real world object have different names, or different real world objects have the same schematic name. Implicit attributes and absent attributes are another issue that has to be contended with, if the task of structural integration is to be successful. Dependency conflicts arise when for example a salary attribute is represented in terms of net salary and tax in another schema. It is the aim of this project to be able to automatically identify and provide solutions for these types of data and schema inconsistencies.

Semantic Integration: Semantics is defined as “the meaning or the interpretation of a word, sentence, or other language form (World’s Information Systems. Semantics, 2004)”. Semantic integration deals with the issues that are raised when data sources have heterogeneous semantics, that is, the meaning of certain data constructs can be ambiguous or hold a different meaning. The designers of the system have, in essence, conceptualized the database in different terms. It may not be possible, based on the information held about the source, to decide whether or not two relations which have the same name represent the same thing. Therefore, if we try to integrate these two objects we may end up integrating relations with heterogeneous semantics and thus we will get incorrect results that make no sense whatsoever (Patrick Ziegler and Klaus R. Dittrich, 1998).

It is therefore necessary to ensure that data that is to be integrated is semantically correct. Semantic integration is defined as “the task of grouping, combining or completing data from different sources by taking into account explicit and precise data semantics in order to avoid that semantically incompatible data is structurally merged (Patrick Ziegler and Klaus R. Dittrich, 1998).” It follows that only data that is deemed to be related to the same real world object must be combined. The problem with this is that there are no set semantic rules that apply for every human user. Semantic heterogeneity can be overcome with the help of ontologies, which are defined “as formal, explicit specifications of a shared conceptualization (Gruber, 1993).” Semantic integration is a field in which much research is being carried out and many open problems still have to be overcome. For this reason, semantic integration will not be covered within the scope of this project.

Mining Models with data Integration Techniques

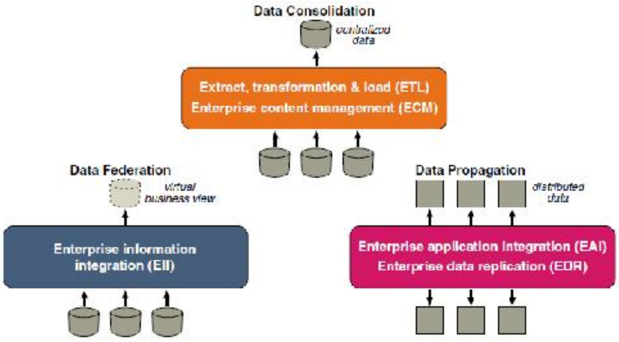
There are several organizational levels on which the integration can be performed. Major techniques involved in data integration are Data consolidation, Data propagation and Data federation.

Data consolidation involves capturing of data from multiple source systems and integrating into a single persistent data store. The latency of the information in the consolidated data store depends upon whether batch or real time data consolidation is being used and how often the updates are being applied to the data store (Chidan and Apte, 2002).

Data propagation involves replicating data in different locations from different sources. Technologies include replication, database log scrapers and change data capture software.

Data Access uses search capabilities to make information accessible to uses via searchable indexes, aggregations and caches using the same type of search technologies that drive

Internet search. The application of search technology in an Enterprise is known as Enterprise Information Access.


Data Consolidation
centralized data
Extract, transformation & load (ETL)
Enterprise content management (ECM)
Data Federation
virtual business view
Enterprise information integration (EII)
Data Propagation
distributed data
Enterprise application integration (EAI)
Enterprise data replication (EDR)

Data consolidation is the main approach used by data warehousing applications to build and maintain an operational data store or an enterprise data warehouse. Data consolidation can also be used to build a dependent data mart, but in this case the consolidation process uses a single data source (i.e., an enterprise data warehouse). In a data warehousing environment, ETL (extract, transform, and load) technology is one of the more common technologies used to support data consolidation (Sheth, Amit and Larson, James, 1990). Another data consolidation technology is ECM (enterprise content management). Most ECM solutions focus on consolidating and managing unstructured data such as documents, reports, and Web pages.

Data federation enables a single unified virtual view of one or more source data files. Data federation technique normally employs a metadata reference file to connect related customer information together based on a common key. Data federation always *pulls* data from source systems on an on-demand basis (Michael Goebel and Le Gruenwald, 1999). Any required data transformation is done as the data is retrieved from the source data files. Enterprise information integration (EII) is an example of a technology that supports a federated approach to data integration. The main advantages of a federated approach are that it provides access to current data and removes the need to consolidate source data into another data store. Data federation, however, is not well suited for retrieving and reconciling large amounts of data or for applications where there are significant data quality problems in the source data. Another consideration is the potential performance impact and overhead of accessing multiple data Sources at run time.

Data federation may be used when the cost of data consolidation outweighs the business benefits it provides. Operational query and reporting is an example where this may be the case. Data federation can be of benefit when data security policies and license restrictions prevent source data being copied. Syndicated data usually falls into this latter category. It can also be used for as a short-term data integration solution following a company merger or acquisition.

Data propagation involves replicating data in different locations from different sources. Technologies include replication, database log scrapers and change data capture software. Data Propagation applications copy data from one location to another. These applications usually operate online and *push* data to the target location; i.e., they are event-driven. Updates to a source system may be propagated asynchronously or synchronously to the target system. Synchronous propagation requires that updates to both source and target systems occur in the same physical transaction. Regardless of the type of synchronization used, propagation guarantees the delivery of the data to the target.

This guarantee is a key distinguishing feature of data propagation. Most synchronous data propagation technologies support a two-way exchange of data between a data source and a data target. Enterprise application integration (EAI) and enterprise data replication (EDR) are examples of technologies that support data propagation. Data propagation implementations vary considerably in both perform and data restructuring and cleansing capabilities. Some enterprise data replication products can support high volume data movement and restructuring, whereas EAI products are often limited in their bulk data movement and data restructuring capabilities. Part of the reason for these differences is that enterprise data replication has a data-centric architecture, whereas EAI is message-or transaction-centric.

Data Integration in the Sirup (Semantic Integration Reflecting User-Specific Semantic Perspective) Approach

Mapping all data to one single domain model causes users to adapt to one single conceptualization of the world. This contrasts to the fact that receivers of integrated data widely differ in their conceptual interpretation of preference for data. They are generally situated in various real-world contexts and have different conceptual models of the world in mind (Goh *et al*., 1994). These models do not only vary between different people in the same domain, but even for the same individual over time (Gaines and Shaw, 1989). COIN (Goh *et al*., 1994) was one of the first research projects to consider the different contexts data providers and data receivers are situated in. In our research, we continue the trend of taking into account user specific aspects in the process of semantic integration. We address the problem how individual mental domain models and personal semantics of concepts can be reflected in data integration to provide tailor-made integration for personal information needs.

In the SIRUP (Semantic Integration Reflecting User-specific semantic Perspectives) approach (Ziegler and Dittrich, 2004), we found how data equipped with explicit, query able semantics can be effectively preintegrated on a conceptual level. Thereby, we aim at enabling users to perform declarative data integration by conceptual modeling of their individual ways to perceive a domain of interest. Origin of our research is the observation that different users often have diverse views of reality i.e., they perceive and conceptualize the same real world part differently, according to their relative points of view, their information needs, and expectations (Kent, 1978). Additionally, none of these co-existing views of the real world can be regarded as being more correct than another because each view is intended for a worthy purpose (Sølvberg, 1997). In general, we refer to this phenomenon as data receiver heterogeneity. Imposing a single global schema for all users can have severe limitations that seriously interfere with the users' individual work because thereby, data receiver sovereignty is violated.

Sovereignty of data receivers refers to the fact that using integrated data must be non-intrusive (Scheuermann *et al*., 1989); i.e., users should not be forced to adapt to any standard concerning structure and semantics of data they desire. Therefore, to take a "one integrated schema fits all" approach is definitely not a satisfactory solution. We generally subsume problems that cause a single global schema to be inappropriate for particular users as perceptual integration mistakes (Ziegler and Dittrich. 2004). These include:

Data selection mistakes are caused when data that is available through the global schema is; from the users' perspective, inappropriately collected and selected from a given data source for example, by only including particular local relations in the global schema.

- Source selection mistakes occur when the decision of the global schema designer, which data sources to incorporate into the global schema, differs from individual users' preferences for data from various origins.
- Entity granularity mistakes refer to the fact that the degree of granularity in which information is represented in the global schema can be too fine-grained (specialized) according to the requirements of individual users for e.g., by integrating a "seminar" and a "colloquium" relation into a general global "course" relation.
- Attribute granularity mistakes are problems of inadequate granularity concerning attributes of entities in the global schema.
- Data semantics mistakes arise when the global schema provides an integrated view on data that is semantically not related according to the individual perception of specific users. For instance, data concerning lectures and seminars may be globally merged since both represent similar forms of teaching. However, this is not useful for people who are only interested in seminars because seminar information is blurred with lectures.

Last, but not least, data taxonomy mistakes occur when generalization / specialization hierarchies given by the global schema do not fit the perspective of the particular domain according to individual users.

Benefits of data Integration in Business Intelligence

While adding data mining scores and predictive models directly in the database is beneficial, there is additional value to be gained by integrating data mining scoring inside the BI platform.

- Business users can view predictive reports in a wide variety of user interfaces.
- Highly formatted predictive reports provide the easiest possible user consumption and professional presentation.
- Personalized messages and predictive reports can be delivered to very large user populations based on alerts or schedules.
- Ad hoc query and analysis that includes predictive metrics is possible without requiring knowledge of SQL, table structures, or predictive models.
- Business analysts can perform further analysis, such as slice-and-dicing data, ad hoc report creation, drilling, pivoting, and sorting, on predictive reports.
- Strict security is applied to users within, and outside the organization.

Future challenges

Data integration in the Business is not a new challenge, but it is a recurring one that has only recently been unfolded as a major challenge in part driven by technology development producing increasing amounts and type of data. However it is become increasingly clear that to be able to integrate across different types of is not only an opportunity but also a competitive advantage within the business Intelligence research community. While the availability of genomics data is reasonably well provided for by publicly accessible and well-maintained data repositories, there is a need for improved annotation standards and requirements in data repositories to enable better integration and reuse of publically available data. The data exploitation aspect of data integration is probably the one that requires most attention, as it involves

- The use of prior knowledge and its efficient storage,
- The development of statistical methods to analyze heterogeneous data sets and
- The creation of data explorative tools that incorporate both useful summary statistics and new visualization tools.

Conclusion

In this paper, we gave an overview of issues and principal approaches in the area of integration seen from a database perspective. The most difficult integration problems are caused by semantic heterogeneity; they are being addressed in current research focusing on applying explicit, formalized data semantics to provide semantics-aware integration solutions. Despite this, considerable work remains to be done for the vision of truly user-specific semantic integration in form of efficient and scalable solutions to become true.

REFERENCES

Chidanand Apte, Bing Liu, Edwin P.D. Pednault, Padhraic Smyth, 2002. "Business Applications of Data Mining," Communications of the ACM, Vol. 45, No. 8.

Gaines, B. R. and Shaw, M. L. G. 1989. Comparing the Conceptual Systems of Experts. In 11th International Joint Conference on Artificial Intelligence (IJCAI 1989), pages 633–638, Detroit, Michigan, USA, August. Morgan Kaufmann.

Goh, C. H., Madnick, S. E. and Siegel, M. Context Interchange: Overcoming the Challenges of Large-Scale Interoperable Database Systems in a Dynamic Environment. In Third International Conference on Information and Knowledge Management (CIKM 1994), pages 337–346, Gaithersburg, USA, November 29 - December 2, 1994. ACM.

Gruber, T. 1993. "A translation approach to portable ontology specifications". In: Knowledge Acquisition. 5: 199-199.

John Clear *et al*.: NonStop SQL/MX Primitives for Knowledge Discovery. *KDD 1999*: 425-429.

Kent. W. 1978. Data and Reality. Basic Assumptions in Data Processing Reconsidered. North-Holland, Amsterdam.

Michael Goebel, Le Gruenwald, 1999. “A Survey Of Data Mining and Knowledge Discovery Software Tools,” SIGKDD Explorations, Vol. 1, Issue 1. Pg 20, ACM SIGKDD.

Patrick Ziegler and Klaus R. Dittrich, “Three Decades of Data Integration-All Problems Solved?”Database Technology Research Group, Department of Informatics, University of Zurich Page 7.

Scheuermann, P., Elmagarmid, A. K., Garcia-Molina, H., Manola, F., McLeod, D., Rosenthal, A., and Templeton, M. 1990. Report on the Workshop on Heterogenous Database Systems held at Northwestern University, Evanston, Illinois, December 11-13, 1989. SIGMOD Record, 19(4):23–31,.

Sheth, Amit P. and Larson, James A. 1990. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Computing Surveys*, 22(3).

Sølvberg. A. 1997. Data and What They Refer to. In Conceptual Modeling, Current Issues and Future Directions, Selected Papers from the Symposium on Conceptual Modeling, Los Angeles, California, USA, held before ER 1997, pages 211–226. Springer.

Surajit Chaudhuri, 1998. Data Mining and Database Systems: Where is the Intersection? Data Engineering Bulletin 21(1).

World’s Information Systems. Semantics." The American Heritage® Dictionary of the English Language, Fourth Edition. Houghton Mifflin Company, 2004. 15th Apr. 2010.

Ziegler, P. and Dittrich, K. R. 2004. User-Specific Semantic Integration of Heterogeneous Data: The SIRUP Approach. In First International IFIP Conference on Semantics of a Networked World (ICSNW 2004), pages 44–64, Paris, France, June 17-19, Springer.

****