International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 3, Number 9 (2013), pp. 971-976

http://www. irphouse.com /ijict.htm

Deriving Business Intelligence from Unstructured Data

Vedika Gupta[1] and Nitasha Rathore[2]

[1,2]CSE Department, HMR Institute of Technology and Management, Hamidpur, Delhi, INDIA.

Abstract

The bulk of an organization's data is unstructured that is the information that doesn't nicely fit into the typical data-processing format. It remains stored, but unanalysed. The information hidden or stored in unstructured data can play a critical role in making decisions, understanding and complying with regulations, and conducting other business functions. Integrating data stored in both structured and unstructured formats can add significant value to an organization. Such integrated data will define the Total Data Warehouse infrastructure so an organization can derive a single version of truth. Analyzing and processing text can reformat this data and merge it with traditional structured data for greater corporate insight. We propose to use text tagging and annotation, to derive business intelligence from business invoices of a company. After the extraction of facts and dimensions from unstructured textual data of invoices, unstructured data is transformed into a relational form and can be stored in a data warehouse. These steps offer businesses an insight into the context, or true meaning, of the unstructured text.

Keywords: Data Warehouse, OLAP (Online Analytical Processing), Unstructured Data, Total Data Warehouse (TDW), Business Intelligence (BI).

1. Introduction

Data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data which helps in decision-making process [1]. According to Kimball [2], a data warehouse is designed for querying and analysing structured data which can be divided into facts and dimensions. Structured data has a pre-defined schema and is

record oriented whereas Unstructured Data (USD) is vast, freeform and exists in variety of forms. It poses difficulty in querying and analysis due to lack of well-defined schema.

The organizations successfully apply the Data Warehouse and OLAP technologies to build decision support systems for organizing and analyzing the huge amounts of structured data that companies store in their databases. Whereas the need of the hour lies in the discovery of such methodologies and tools that can deal with the massive storage and retrieval of documents with large text-rich sections and hence cater to BI applications. Companies and enterprises also circulate an enormous amount of information as text-rich documents – pdf, word files, e-mails, chat files, blogs, organization forums and many other. Nowadays, World Wide Web has become the greatest source of information, organizations can now find highly valuable information about their business environment on the Internet, which is a benefit although but has created around 80% of the data floating to be in unstructured format which is difficult to analyse and store in data warehouse. Data warehouses are the huge data repositories which stores historical and current data of enterprise worlds and thus keep all the data at one place. Structured data is stored easily into the data warehouse but unstructured data poses problem in such storage. But actionable knowledge is pertinent in unstructured textual documents. The need to manage unstructured data arises due to the fact that more than three-fourth of information on internet is unstructured. The advantages one can get out of Unstructured Data management is Business Value, Better information, Timely information, Relevant Information.[3]

Traditional unstructured data sources are very high in volume. So, the challenges facing data are as follows [4]: getting the right information from it, transforming it into knowledge, analyzing it to find patterns & trends, storing information for fast & efficient access, managing the workflow and finally, making useful BI reports.

To investigate any fact or incident, we need to analyze multi format data from multiple sources in different time frames. The integrated information architecture facilitates better insight of multi-dimensional information for the targeted entities. It also provides better insight, more powerful statistical, semantic, co-relational and reporting capabilities. Faster read and write ability provides collected data in near real time analytical capabilities. The requirement is to make the warehouse capable of handling large data sets that are challenging to store, analyze, search, visualize, share, and manage.

2. Total Data Warehouse (TDW) using Text Annotation

The process of capturing intelligent information from unstructured data is performed in two phases. In the first phase, structure is added to the unstructured data via named entity extraction. After that results are integrated with structured data. The output obtained from phase 1(that will be TDW), acts as an input to phase 2, in which BI application requirements are catered by the TDW.

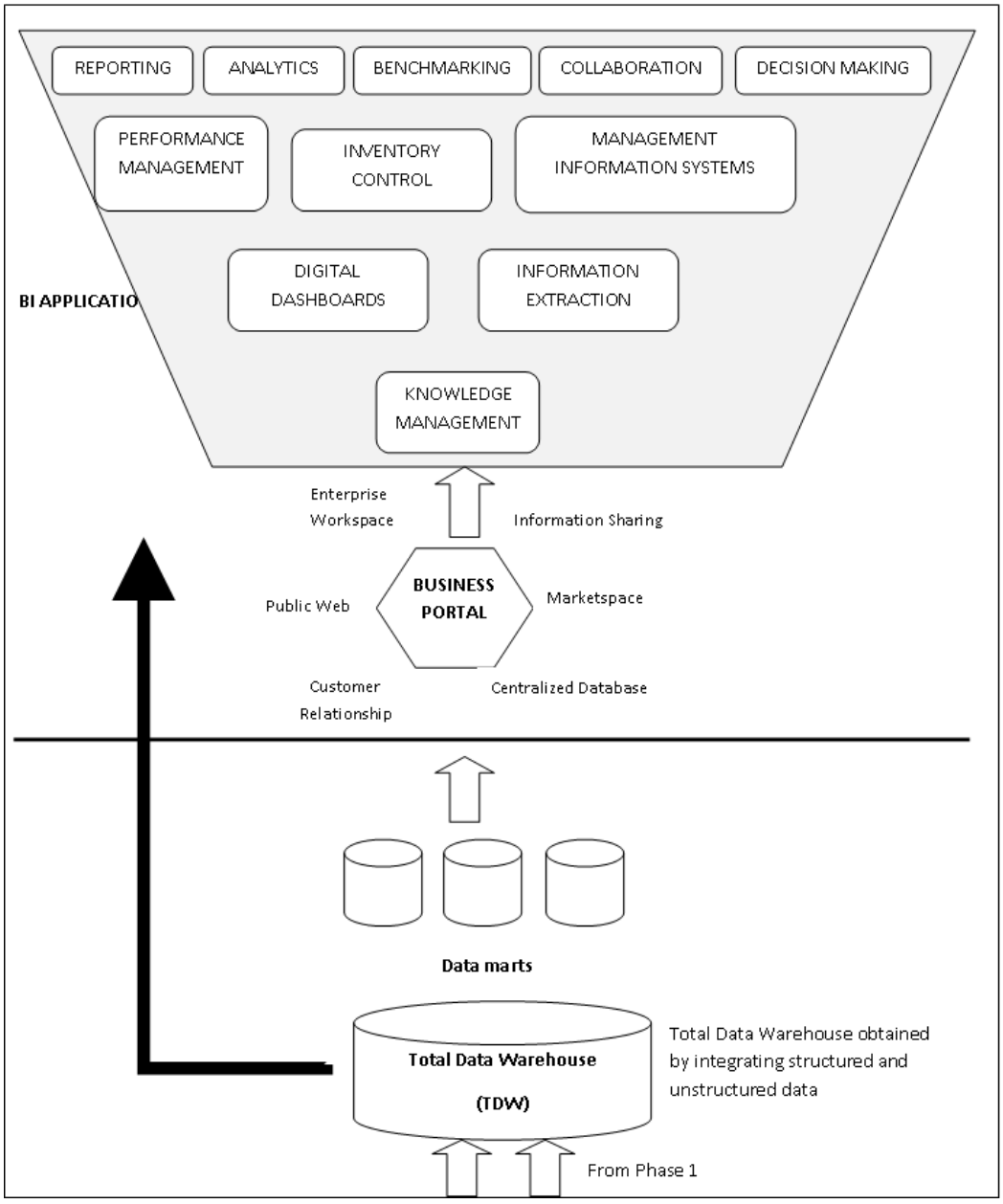

Figure 1: ETL, text tagging, and annotation are used to build the total data warehouse (phase 1).

As shown in Fig.1, data within an enterprise can come from traditional transactional sources such as an RDBMS, legacy systems, and repositories of enterprise applications, and from unstructured data sources such as file systems, document and content management systems, and mail systems.

To build an effective decision-support backbone, this data must be moved into the TDW. An ETL process executes the required formatting, cleansing, and modification before moving data from transactional systems to the TDW. In the case of unstructured data sources, the tagging and annotation platform extracts information based on domain ontology into an XML database. As in Fig.2, extraction of data from an XML database into the TDW is accomplished with an ETL tool. This materializes the unified data creation into the TDW—the foundation for the organization's decision-support and BI needs.

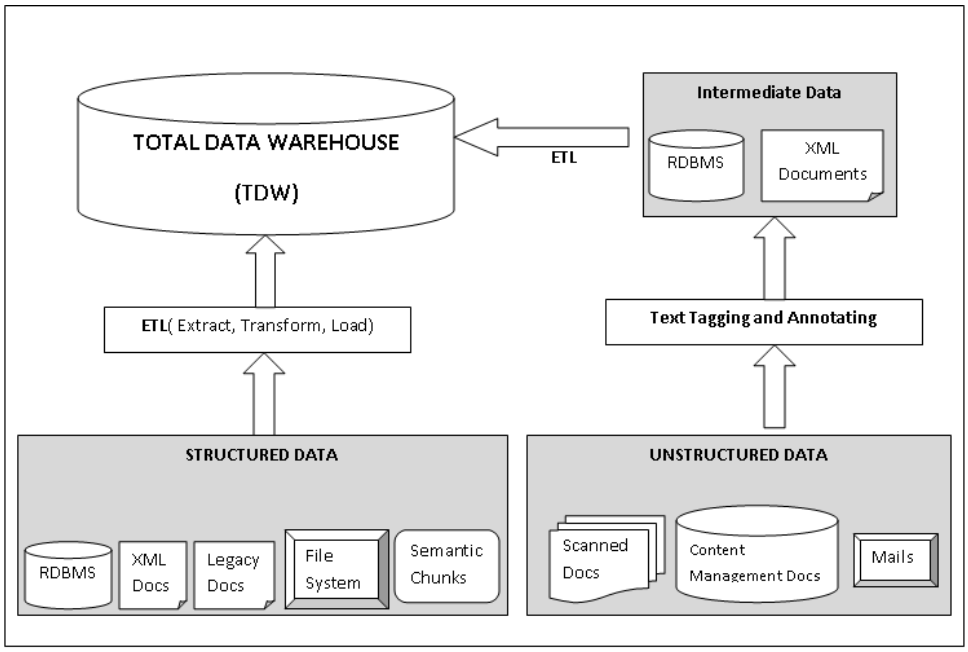

Figure 2: Building business intelligence applications from the TDW: phase 2.

3. Product Performance Insights from Customer Warranty Claims Data

An example of BI application is to analyse warranty claims data in case of a product's failure or defect. Warranty claims have some structured and some unstructured content. These claim forms constitute warranty data that is to be analysed for gaining business intelligence, and moreover diagnosing the problem in the company product. A claim form has details such as product id, product name, model number, date and time of purchase, customer name and address , defects encountered etc. These forms are appended into a database (may be as BLOB). Some of these entered information is structured data which is in defined format and has finite answers in defined fields. But the comments section in the form has freeform English language text which is

unstructured and the most important information for understanding and analyzing the defect encountered. The huge number of claim forms renders the manual reading of all comments time-consuming and practically difficult. The idea is to automate text analysis that collects claim form information by extracting information from unstructured data and linking it with an external knowledge base.

Fig.3 illustrates a claim form entered by a customer and received by a company repair center. The form is partially structured for the reason that some fields have a defined format. Other fields allow the user to enter paragraph form text. The user provides details that describe the technical defect in the product. The text of the user's comments contains many domain specific entities. For example, camera is an entity of type "Computer Parts", crashed is a "Defect" entity. Similarly technician's problem analysis is also written in a natural language text, from which many real world business entities can be derived. The basic technology used to tag and annotate the text can be based on a dictionary lookup created from an external knowledge base.

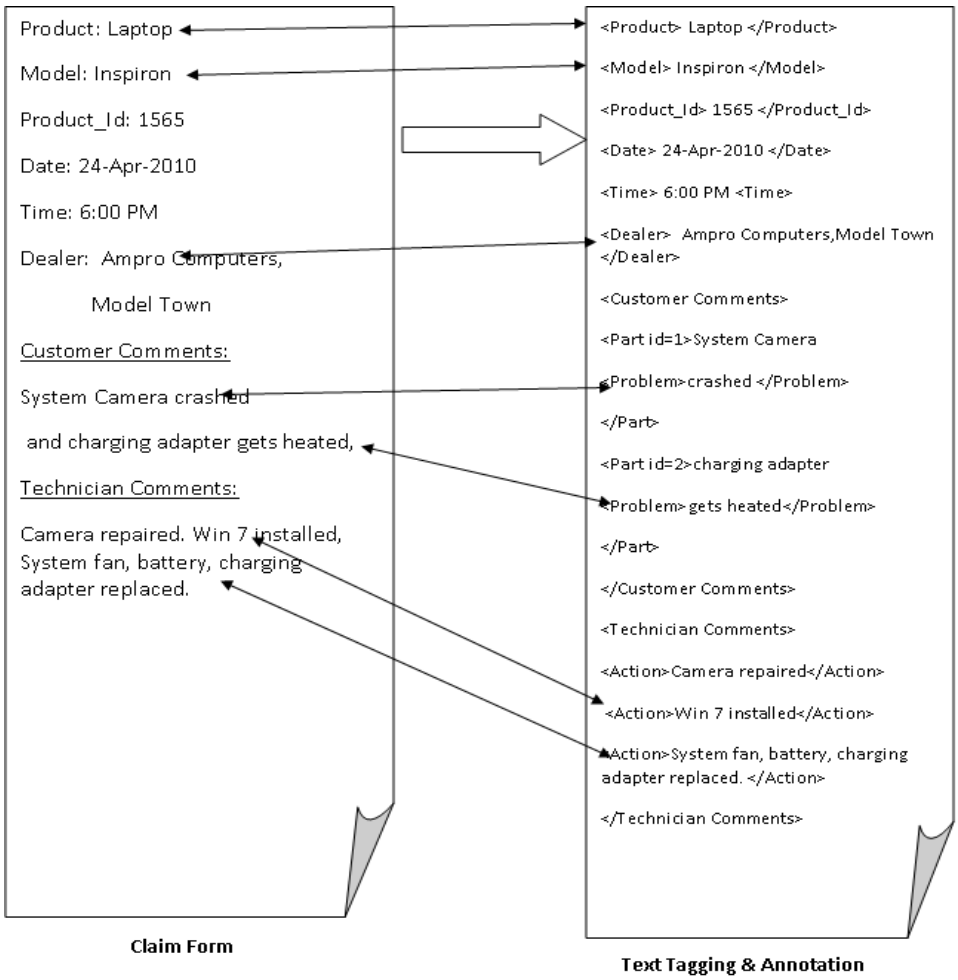

Figure 3: Annotating warranty claims data.

As depicted on the right side of Fig.3, the output of the text tagging and annotation process is in XML file format containing the extracted entities enclosed within XML tags. The XML file produced by the text tagging process is in a form amenable to query, search, and integration with other structured data sources.

The first step in gathering business intelligence from these claim forms is to tag and annotate the text—this is the named entity extraction. In the second step, the tagged data is combined and analyzed with an external structured data repository. The output of text tagging can also be imported into a relational database.

4. Summary & Conclusion

Text tagging and text annotation plays efficient role to integrate structured and unstructured data. The resulting total data warehouse becomes the basic framework for BI applications. The BI benefits of deploying a text analytics methodology speeds up the identification of product defects and their consequent repairmen or replacement by integrating the unstructured data entered by customers and technical assistant with the structured data stored in a relational database. As shown, the methodology can be applied to help an enterprise gather intelligent information from meetings text or gain intelligence about product performances from customer warranty claim forms. The total data warehouse can be employed as a framework for efficient and accurate business decisions

References

[1] W.H. Inmon. *Building the Data Warehouse*. John Wiley, 1993.

[2] Kimball, Ralph; Margy Ross, Warren Thornthwaite, Joy Mundy, Bob Becker (2008). *The Data Warehouse Lifecycle Toolkit* (2nd ed.). Wiley. ISBN 978-0-470-14977-5.

[3] Ayaz Ahmed Shariff K, Mohammed Ali Hussain, Sambath Kumar Sneddon, “Leveraging Unstructured Data into Intelligent Information- Analysis and Evaluation”, 2011 International Conference on Information and Network Technology IACSIT Press, Singapore IPCSIT vol.4 (2011)

[4] Vedika Gupta, Anjana Gosain. “Tagging Facts and Dimensions in Unstructured Data” International Conference on Electrical, Electronics & Computer Science Engineering (EECS) May 2013.