International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 7, July - 2013

Extracting Facts And Dimensions From Unstructured Data For Business Intelligence

Vedika Gupta
HMR Institute of Technology and Management
Delhi

Abstract

Structured data and unstructured data are handled as two distinct information entities. This often results in failure of decision management as information embedded in unstructured data (USD) can play a vital role in making business decisions for the fact being that around 80% of information resides in unstructured format in the organizations. So there is a need of a framework that meaningfully relates and integrates structured and unstructured data and would act as a total data warehouse (TDW) that may serve as the foundation business intelligence. This way data would be treated purely as a bunch of information for gaining business insight, irrespective of the structure of the data. In this paper, we intend to propose hypothesis to extract structure (facts & dimensions) from the unstructured textual data. The motive is to find analysis issues (i.e., facts), or viewpoints (i.e., analysis dimensions).

1. Introduction

The amount of data stored in an enterprise is growing quickly. The ability to access and analyse data sources for intelligent decision making is a key element to an organization's success. Enterprises evolve and transform over time, resulting in a heterogeneous world of information where data is distributed across a variety of sources and systems. Data stored in different systems, locations, formats, and schemas poses a challenge to integration and usability. Enterprises are increasingly interested in accessing unstructured data and integrating it with structured data.

The single integrated view of the truth can only be inferred after correlating relevant facts from continuous textual data stream. Traditional textual data sources are rich in volume & variety- emails, chats, IM, documents, spreadsheets, letters, memos, thesis work, pdf, word files, text files, minutes of meetings, sms, financial transactions, travel information, internet usage reports etc. In order to investigate any relevant fact or incident from these unstructured textual sources, multiple dimensions of data from multiple data sources need to be analysed along different time frames. Therefore, it is important to have an integrated information architecture that facilitates better insight of multi-dimensional information to cater business decision and important events.

Data warehouse is defined as a subject-oriented, integrated, time-variant, and non-volatile repository of data which assists decision makers and business analysts in decision-making process [1]. 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. Unstructured data consists of freeform text containing natural language. The fundamental difference between structured data and unstructured data is that structured data is organized in a highly mechanized and manageable way. Structured data is ready for seamless integration into a database or well structured file format such as XML. Unstructured data, by contrast, is raw and unorganized.

The pertinent information stored in unstructured data can play a critical role in making decisions, understanding potential risks 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 our Total Data Warehouse (TDW) framework so an organization can derive a single version of business-wise significant facts.

Text tagging and annotation consists of analysing freeform text and identifying terms (for example, proper nouns and numerical expressions) corresponding to domain-specific entities.

As data is continuously collected and created, companies have difficulty just storing it, missing any opportunity to leverage the information. The wave of unstructured data has the potential to flip the burden of data management into the opportunity of new value

creation. Yesterday's solutions don't accomplish this today and will be even less effective tomorrow. Unstructured content lacks in any form or structure we commonly associate with traditional corporate data. The text can be in any language; it may follow or not follow grammatical rules, or may be amalgamation of words and numbers. But the big difference between unstructured data and structured content, as I see it, is not what it is but what you do with it. While the volume of data has grown exponentially over the last few decades, the fundamental and underlying technology on which we store data hasn't. There had been improvements in densities (volumes of data) and connectivity (to provide accessibility to data), but the pace of data growth has overwhelmed the benefits of these technological advancements.

A data warehouse is designed for querying and analysing structured data which can be divided into facts and dimensions [2]. The foundation of deriving facts, measures and dimensions is not inherently set into the relational model or tied to structured data. Unstructured data also contains facts and dimensions that provide insight into business intelligence applications. Any unstructured data source has a rich set of dimensions, even if they are only implicitly attached to the actual data structures. The basic integration approach is to actually tie the respective dimension keys to each data source [6].

Textual information is needed to specify the reason for the existence of numerical facts, and this reason helps in charting out Business Strategies and Decisions intelligently. Analysing unstructured data allows users to find patterns hidden deep inside large data sets. Content management, business intelligence, knowledge management and similar applications increase the need to integrate rather than append unstructured data. The "BLOB" approach works to append but not to integrate. A textual file can be stored as a BLOB into the database. But, that activity will be futile. The resulting data base would not be actionably processed for Business Intelligence. For the sake of producing meaningful results, the text must be transformed in such a way that only the incomprehensible pertinent concepts come into the forefront lucidly. In a typical 'minutes of meeting' document, there may be much more information than that needed for the specific issue. We try to identify and filter pertinent information from that textual source. In this paper, we would be analysing text to assort the facts into their relative dimensions.

The layout of the paper is as follows. In section II, we discuss the related work existing approaches utilizing XML for handling unstructured data in data warehouse. In Section III, we present our methodology adopted to annotate facts and dimensions in a typical 'minutes of meeting'. Lastly, we conclude in section IV. References to this work are given in section V.

2. Related Work

USD accounts for many useful facts that are important and sometimes quintessential to be analysed by analysts and business owners to make intelligent decision. Online analytical processing (OLAP) systems have been effective for analysing and mining structured numeric data, but they face challenges in handling text data. Since there is a fast growing need to handle varied unstructured data floating in a warehouse so that more business information can be exploited and derived from the hidden facts in textual data in a document warehouse, various authors have kept forward various hypotheses to extract facts and dimensions from the unstructured content. In order to extract structured information from unstructured data sources, multi-dimensional analysis of the content forms the core of many of the existing approaches to derive facts and dimensions. XML is yet another eminent solution in this regard. Research is going on in this direction. Various authors have taken the XML way to integrate unstructured format into structured format. Wrangling data can be captured in a semi-structured way using XML, which involves tagging the data and then managing the data using the tags. This auxiliary business data can be extremely important to understand because it holds the key to improving business conversion rates, to achieving higher levels of customer satisfaction and to increasing the competitiveness of the business.

In [3] Authors have modified an existing multi-dimensional data model and algebra for representing the R-cubes. The framework proposed develops a contextualized XML document warehouse which integrates the traditional corporate data warehouse with and a document warehouse. This contextualized warehouse contains facts and their contexts - represented as R-cube, characterized by two dimensions namely: the relevance and the context dimensions.

Authors have used XML as Information Retrieval technique.

In [4] authors have discussed about Text Analytics to deal with unstructured & presented a framework for handling unstructured data that are in documents so that they fit into business applications like Data Warehouses for further analysis. Text tagging and text annotation have been used as a part of text analytics

[5] Takes the approach of mapping of structured and unstructured data via multi-layer schema. Another

generic XML schema is used for integrating various types of unstructured documents using linguistic matching mechanism, WordNet which identifies semantically related data to enable the integration between both data sources.

In [6], we have taken a typical 'Minutes of Meeting' text, and tagged the facts in the text with the help of XML tags. We have analysed the text further on the basis of some typical 'stop-words' encountered in the text.

3. Business Intelligence from Minutes of Meeting

Access to ideas and opinions of all the committee members, repercussions of the discussions carried out in the meeting are crucial for the progress of an organisation. Decision makers must captivate and interpret a huge amount of information generated every day. However simple the structure of a business text may seem to be, it is non-trivial to fire a query directly on this natural language text in these textual documents to decipher the meaning it beholds. So the basic proposition is to have a framework i.e. a Total Data Warehouse which will store the relational databases and XML documents that will be derived from unstructured text.

The TDW offers an all-encompassing view of an organization's data assets for building BI and decision-support applications. As illustrated in Figure 2, the TDW enables the building of reliable BI and decision-support applications based on comprehensive and one single integrated version of enterprise data. With the TDW, applications such as corporate performance management can produce highly reliable results.

Figure 1 depicts a scenario of building BI applications with a TDW framework.

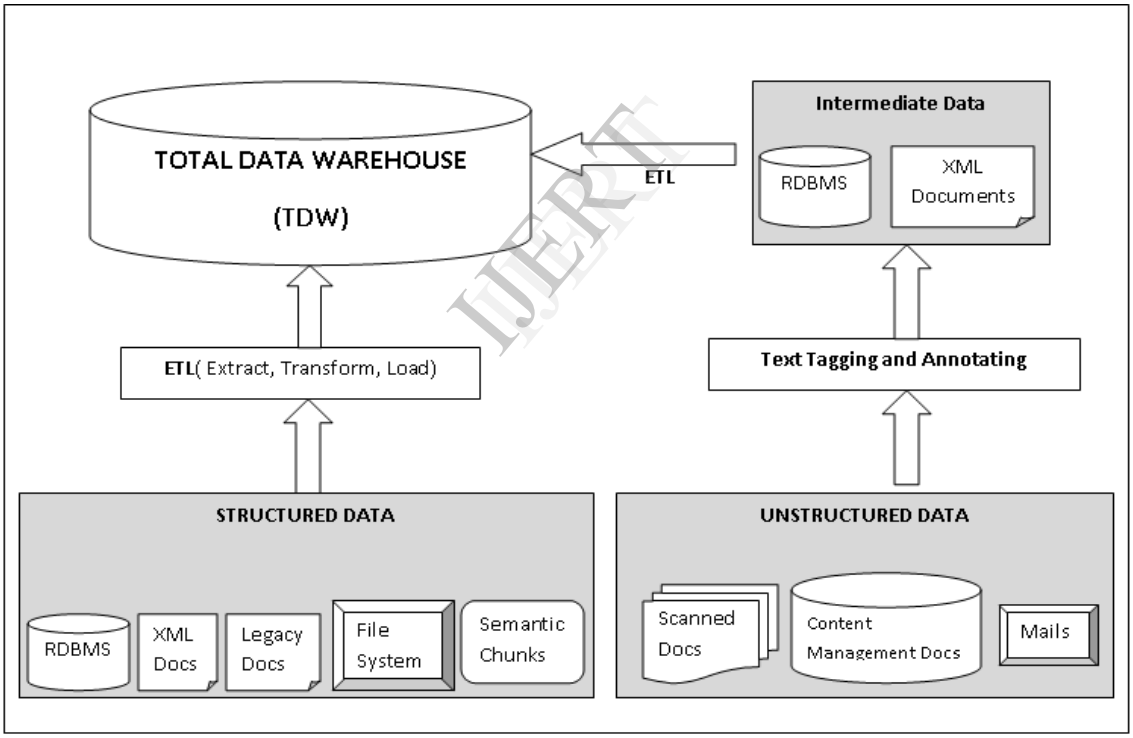

Figure 1. Building business intelligence applications from the TDW

We have taken a piece of text on 'minutes of meeting' as shown in Figure 2. Now from this piece of text we will take out the general information represented as context and measures in the unstructured format by tagging facts with the help of XML tags. The approach is to convert the source text into atomic elements to recognize its underlying structure. These atomic elements in the resulting XML schema are the indivisible pieces of information. Tagging isolated atomic elements in XML is worth to save future confusion, time and space to just reference the relevant text by using proper terminology.

The meeting text is:

The first meeting of the Commission was held at 10.00 A.M. on 23rd March, 2000 in the Committee Room No.1 of India International Centre to discuss agenda items as circulated to the members. The meeting was chaired by Hon'ble Justice Shri M.N. Venkatachaliah. There were 8 agenda items that were discussed.

The entity types present in this text are Subject, Date, Time, Venue, Chair, Agenda. As shown in Figure 1, a text annotator identifying the entities and tagging them. The output can be either semi-structured in the form of XML document or structured in the form of database table. Tagging important and pertinent named entities assists to analyse entity links and relationships. After the extraction of facts and context from unstructured textual data, unstructured data gets transformed into a relational format and stores it in a data warehouse. These steps offer businesses an insight into the context, or true meaning, of the unstructured text.

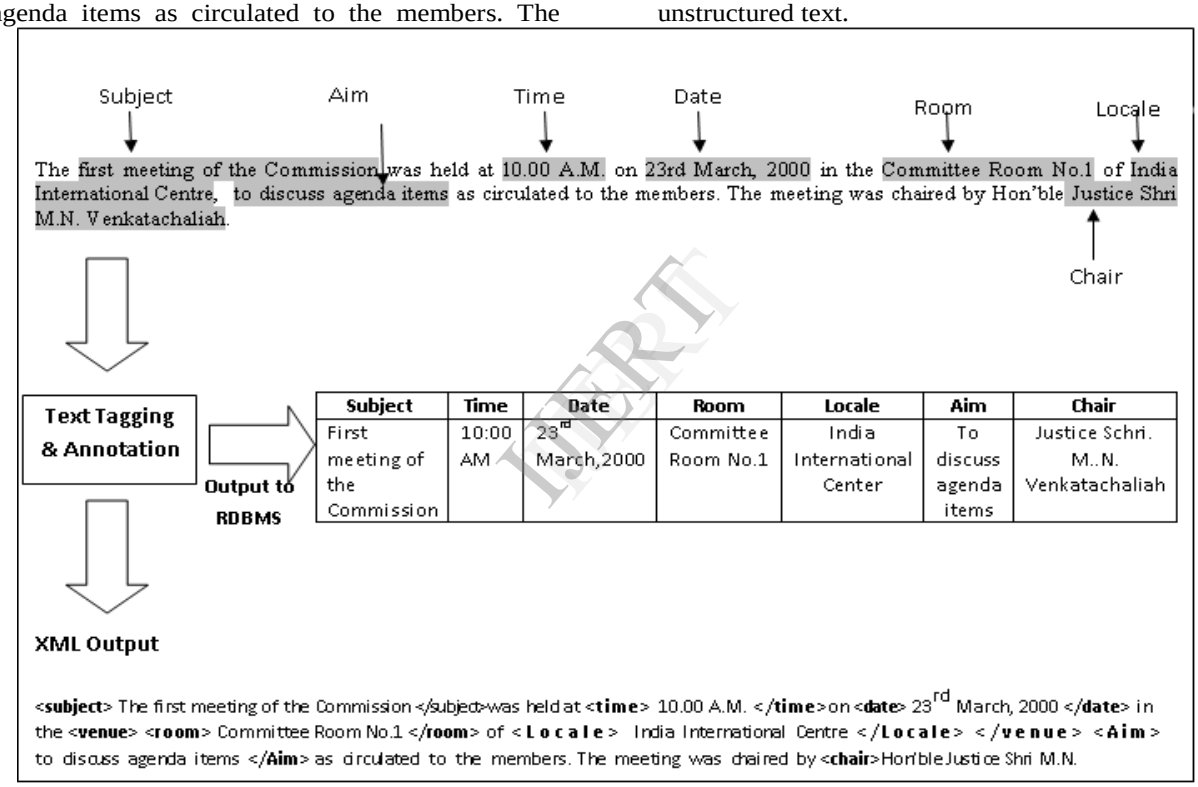

Figure 2. Text annotation and tagging on 'Minutes of Meeting'

Figure 3 depicts potential benefits of integrating the structured and unstructured data. The answer to some queries as listed in Figure 3 cannot be elicited from a single business document. For example, to provide an answer to the query, "List meetings chaired by Justice Shri M.N. Venkatachaliah in the second and third quarter of 2000" one needs to examine a large amount of unstructured content from minutes of meeting documents and discern the relationships between Justice Shri M.N. Venkatachaliah and the meetings he chaired and the meeting dates. The meeting minutes may not directly mention Q2 or Quarter two as the date or time of the event. The knowledge that the months April through June comprise the second quarter of a year comes from an external knowledge base or previous learning.

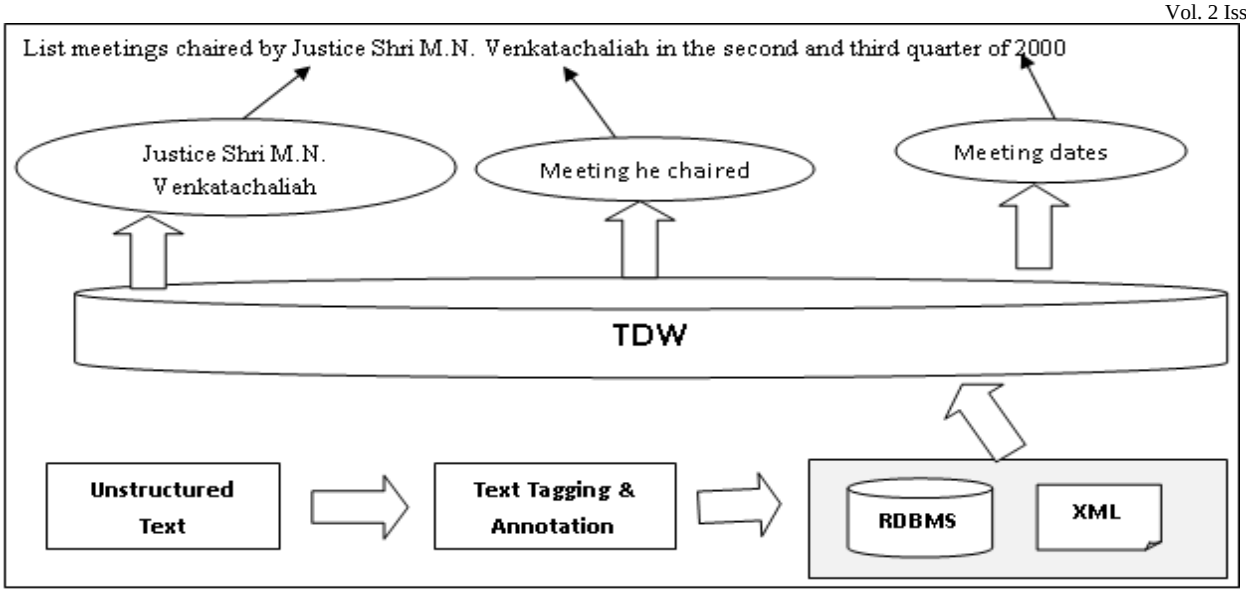

Figure 3. Integrating structured and unstructured data to captivate business intelligence

In another query, a user wishes to know the number of participants in ever meeting of the commission and the participant who attended more than 3 meetings. The meeting minutes does not mention explicitly the name of participant and the number of meetings attended by him. If 3 texts on minutes of meetings mentions the name of same participant, then that person is included in the result of the above query. After putting the relevant textual information into xml format, we can skip out on remaining textual data which does not add much relevance to the context of data.

This is how our XML methodology grasps the implications, relationships and domain of day-to-day language and extracts who did what, when (at what time, date), where (at what place) and under what conditions (circumstances that hold prior to the implication) and the repercussion. It then creates output that is in XML format fused with existing structured data in the content so that it can be mapped to relational tables in data warehouse so that business intelligence (BI) applications can access and put into operation.

4. Conclusion

Unstructured textual data is placed into the data warehouse as text files or BLOBS. These files contain a rich set of facts and dimensions which are otherwise not noticed due to lack of their visibility in a structured format. Therefore, it is required to tag and annotate the facts inherent in the text and its relative dimensions, so that the structures derived from it might be used for knowledge management and business intelligence. In this paper we have taken 'Minutes of Meeting' as an unstructured piece of text and presented a methodology based on XML to tag and locate facts that need to be relationally linked to their dimensions in order to ease BI decision making process. Our solution extracts unstructured data from text documents, transforms it into XML, maps to a relational table and then feeds the data to the data Warehouse, wherefrom the information can be retrieved and analysed using the structured data approaches.

5. 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] J. M. Perez, R. Berlanga, M. J. Aramburu, T. B. Pedersen, "A relevance-extended multi-dimensional model for a data warehouse contextualized with documents," Proc. 8th ACM Intl. Workshop Data Warehousing and OLAP, 2005, pp. 19 - 28.

[4] Kalli Srinivasa Nageswara Prasad, Prof. S. Ramakrishna, Text Analytics to Data Warehousing. (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No.06, 2010, Pg 2201-2207

[5] Ahmad Abdullah Alqarni, Eric Pardede. Integration of Data Warehouse and Unstructured Business Documents, 2012 15th International Conference on Network-Based Information Systems

[6] Vedika Gupta, Anjana Gosain. Tagging Facts and Dimensions in Unstructured Data, International Conference on Electrical, Electronics & Computer Science Engineering, May 2013, Interscience Research Network.