Text Analytics to Data Warehousing

Kalli Srinivasa Nageswara Prasad
Research Scholar in Computer Science
Sri Venkateswara University, Tirupati
Andhra Pradesh , India

Prof. S. Ramakrishna
Department of Computer Science
Sri Venkateswara University, Tirupati
Andhra Pradesh , India

Abstract — **Information hidden or stored in unstructured data can play a critical role in making decisions, understanding and conducting other business functions. Integrating data stored in both structured and unstructured formats can add significant value to an organization. With the extent of development happening in Text Mining and technologies to deal with unstructured and semi structured data like XML and MML(Mining Markup Language) to extract and analyze data, text analytics has evolved to handle unstructured data to helps unlock and predict business results via Business Intelligence and Data Warehousing. Text mining involves dealing with texts in documents and discovering hidden patterns, but Text Analytics enhances Information Retrieval in form of search and enabling clustering of results and more over Text Analytics is text mining and visualization. In this paper we would discuss on handling unstructured data that are in documents so that they fit into business applications like Data Warehouses for further analysis and it helps in the framework we have used for the solution.**

Keywords — Information Extraction (IE); Entity; Semantics; Natural Language Processing (NLP); Parsing.

I.INTRODUCTION

A huge amount of electronic information is either recorded in databases as transactions or available in forms of texts. This information is potentially very valuable to 'companies', 'decision makers', 'partners' and 'competitors' to drive their business. It is easy to derive relations from structured sources for decision making, but the important task for them is to extract relevant information from unstructured data sources for decision making or how to deal with such kind of data. There are available tools and techniques to handle structured data formats, but when it comes to unstructured it is difficult for organizations to process or find relations.

Our paper develops on text analytics in dealing with unstructured data for decision making. We have studied text analytics and its ability to transform textual sources to support structured environments to come up with a framework to deal with unstructured sources. In this paper we will go through different text analytics techniques and drawbacks of text analytics techniques in dealing with "unstructured sources" and we would also discuss data modeling in unstructured cases.

II. UNSTRUCTURED DATA CHALLENGE

Structured data format are usually record oriented, transactions are stored as per predefined data model, which makes it easy to query, analyze, and integrate with other structured data sources. Structured data models are in forms of tables and have relations among model and it is easy to create reports out of it. However unstructured data is contrary to structured sources because of freeform text makes it more difficult to query, search, and extract, and even complicates integration with other data sources.

Traditional unstructured sources look "Versadial develops easy to use telephone recording software, cost-saving call recording kits and turn-key systems that can record up to 256 analog, digital, VoIP or radio channels at once in a single PC". Need to be processed and the relations should be understood and unlocked for better decision making. Today, companies strongly rely on a relational data or transactional data for decision making or for business analysis. Data that is coming from Unstructured sources like emails, conversations and texts are not analyzed leading to business risk.

Text mining has been one of the technique to unlock relations in textual data, even search is termed to find details but decision makers need results rather than links, Since analysts want facts, answers to questions, textual sources should be tamed as per structured sources and the important is to enable decision making support for unstructured data through finding relations and provide knowledge. Text analytics is the answer to overcome "Unstructured data Challenge".

III.TEXT ANALYTICS HOW IT WORKS AND ITS USAGE

In simple terms Text Mining is Data Mining from textual sources and Knowledge discovery from in text, but what makes text analytics different is it looks for structure that is inherent in the textual source materials and either applies linguistic and/or statistical techniques to extract concepts and patterns that can be applied to categorize and classify

uments, audio, video, images. It also transforms "unstructured" information in to data for application of traditional analysis techniques and eventually helps in unlocking meaning and relationships in large volumes of information that were previously unprocessable by computer. This makes Text analytics different and usually termed as superset of text mining.
Typical steps in text analytics include—

"Retrieve documents for analysis"

Apply Statistical and /or linguistic and /or structural techniques to identify, tag, and extract entities, concepts, relationships, and events (features) within document sets.

Apply statistical pattern-matching and similarity techniques to classify documents and organize extracted features according to a specified or generated categorization/taxonomy" [Grimes, Seth2008].

In this paper the solution we utilize text analytics techniques to support unstructured sources for structure formats, for content analysis key is extracting information. Entities and features are like dimensions in a standard decision support model. Text Analytics technique utilizes Information Retrieval and extraction and also use Natural Language Processing for processing textual sources into models that support structured sources.

IV. FRAME WORK

In this solution, we recommend enabling decision support for "Unstructured data", for this we would like to demonstrate the value of text tagging and annotation, a text analytic technique as a preprocessing step toward integrating structured and unstructured data.

Text tagging is a popular technique based on natural language processing and important component of a document processing and information extraction system. Text tagging(Wrapping XML tags) and annotation consists of analyzing freeform text and identifying terms (for example, proper noun and numerical expressions) corresponding to domain-specific entities [Surekha, Asish et al, 2005].

In the figure I, you can see applying tagging technique as part of text analysis for converting plain text to a model that can be processed into the databases or data warehouses. The reason behind using text tagging is because it supports decision making environments when compared to other text analytics techniques like semantics or statistical analysis.

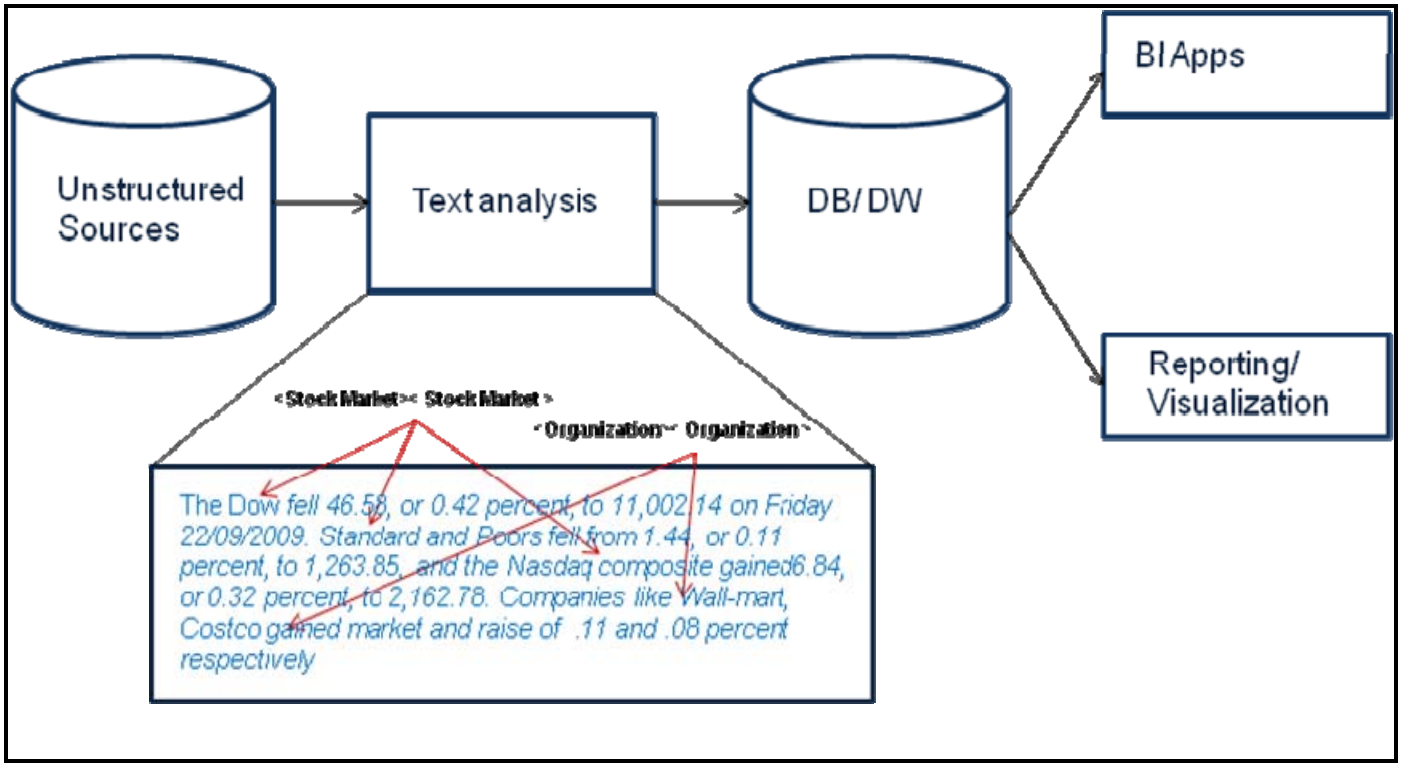


Fig I- Framework to support Text Analytics to Data Warehousing

The drawbacks like shallow parsing and statistical analysis can be enough, for instance, to support classification or categorization just like text mining does, but since our focus is on making unstructured data for decision supporting we need more, and statistical technique can give you rank of a term that is more available in data sets, it can help you get at meaning, for instance by studying co-occurrence of terms but not a data model for further processing.

Consider like text to be processed using our framework :

" The Dow fell 46.58, or 0.42 percent, to 11,002.14 on Friday 22/09/2009. Standard and poors fell from 1.44, or 0.11 percent, to 1.263.85, and the Nasdaq composite gained 6.84, or 0.32 percent, to 2,162.78. Companies like Wall-mart, Costco gained market and raise of 0.11 and 0.08 percent respectively"

Syntactical analysis for the above text would be as follows, syntactical analysis is used for classification, http://www.connexor.eu/technology/machinese/demo/syntax/

In the figureII you can see, how the words in the text sources are split, even though certain techniques helps us segregating words to perform classification or categorization or we use can ranking to find how many terms usually get repeated in document sets that are being processed.

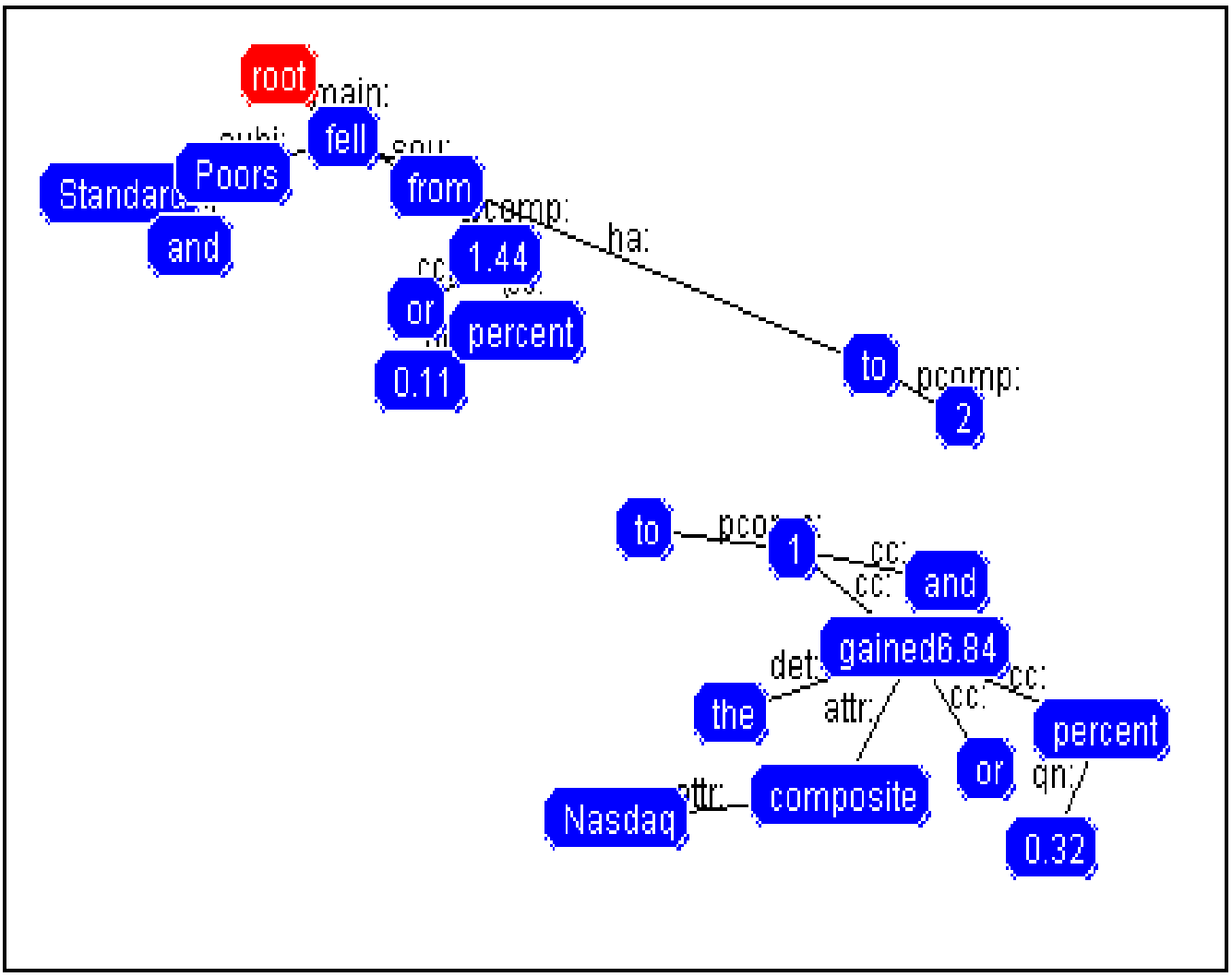


Figure II – Syntctical Analysis model for textual sources.

V. TEXT TAGGING AND ANNOTATION

Text tagging and annotation, also called named entity extraction, forms an important component of many language-processing tasks, including text mining, information extraction, and information retrieval.

Named entity extraction consists of identifying the names of entities in freeform or unstructured text. Among the common types of entities are proper nouns, names, products, organizations, locations, e-mail addresses, vehicle data, times and dates, and numerical data such as measurements, percentages, and monetary values.

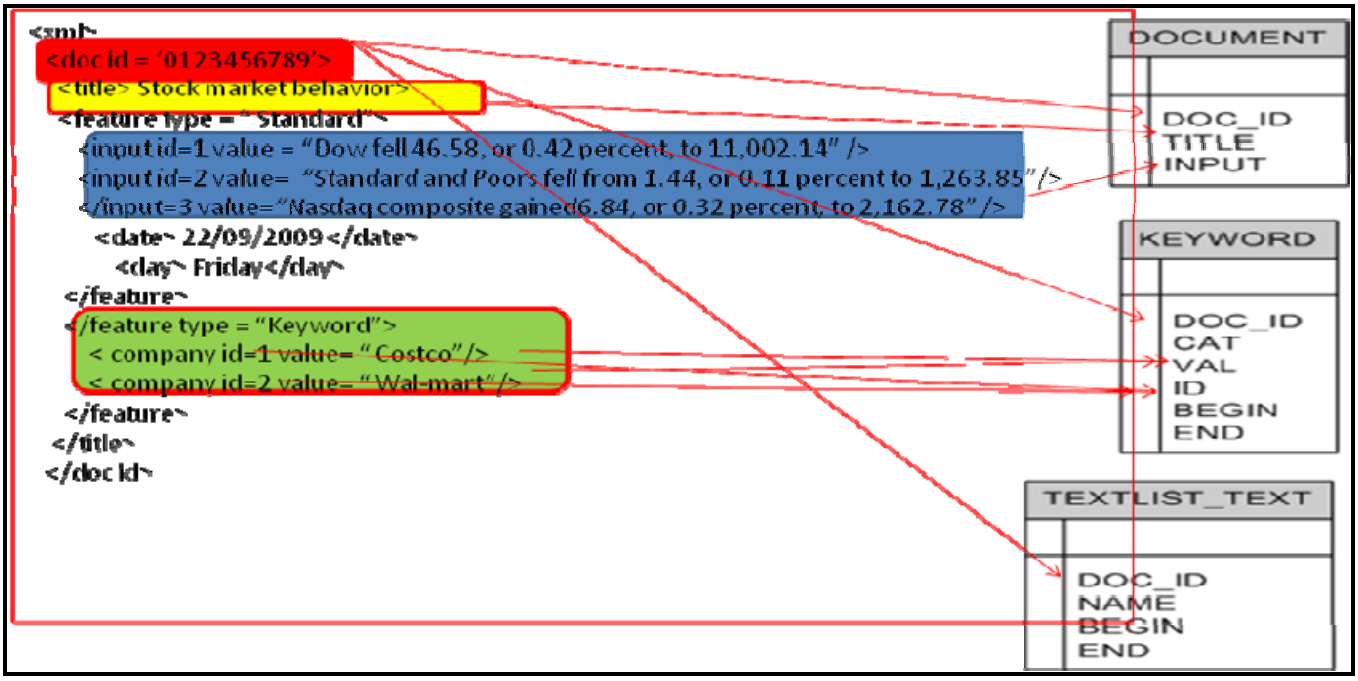
Domain-specific entities are included as well. Named entity extraction has applications in diverse domains, such as detecting chemical and protein names from medical literature; gaining market intelligence by detecting personal names, locations, organization names, and product names in newswire text; finding names of weapons, facilities, and terrorist organizations for military and defence purposes; or building a semantic search applications to overcome the limitation of regular keyword-based search engines.

Several approaches and techniques have been developed to performs named entity extraction, from manually developing a set of rules and using a dictionary or a list lookup from pre-existing databases to linguistic analysis and machine learning.

VI. TAGGING IN ACTION – DATA MODELLING

This section describes the XML(Extreme Markup Language) how tags are written based in the textual sources that are extracted and how structured and unstructured data can be mapped for deriving data model.

As you see in Figure III, each document is represented by an ID number, a title, the original content of the document (“input”), and in form of sets like standard features and keyword features. The standard features represent general structured information from each document, such as date, origin, subject, or country. There is one <kw> element for each such data item. The val attribute of the <kw> element contains the actual value, the cat attribute contains its category, such as “date” or “country”. The keyword features describe the occurrences of words and phrases in the unstructured text content. There is one <kw> element for each distinct word or phrase, each with a value and a category, such as “general noun”, “proper noun”, “verb”, or “phrase”.



Fure III-Tagging and Mapping

VII. STAR SCHEMA

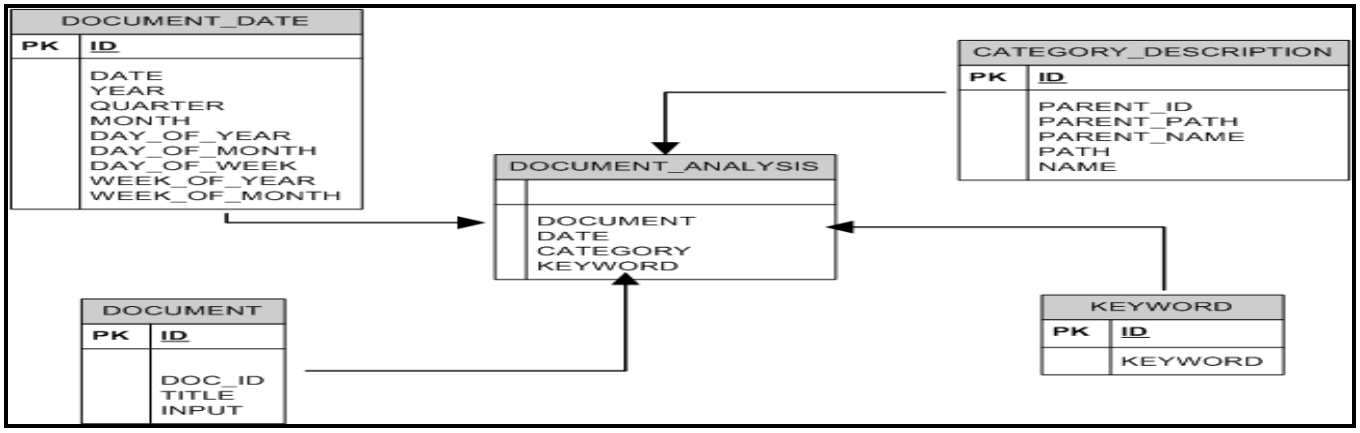


Figure IV- Star Schema

This figure IV shows another possible schema for intermediate output data that is generated from text analytics. This is a star schema consisting of a fact table and four dimension tables. The fact table DOCUMENT_ANALYSIS contains one row for each occurrence of a keyword in a document. The dimension tables describe each analysis entry
in the fact table : the actual date, the keyword values and categories as well as corresponding document title and content. This is a more normalized schema than the four-table schema option, so that repeated storage of identical keywords and categories is avoided.

VIII. HOW XML IS STORED AND PROCESSED

Figure V shows shredding the process of mapping XML elements and attributes into relational tables and columns. One way to shred in is through the use of an annotated XML schema. If the XML data contains an XML schema, it is the easiest and fastest way to perform decomposition. If the mapping is significantly complex and involves multiple tables, existing tools automate both the mapping and decomposition steps. Another, perhaps less-known, method for shredding is through the use of the SQL/XML function XMLTABLE. It is useful when an XML schema does not exist.

The XML annotated form which is the output text analytics can be stored in databases as per the above mentioned process. With the development of XML storage it is possible to store or insert data that is coming from XML documents inside the databases.

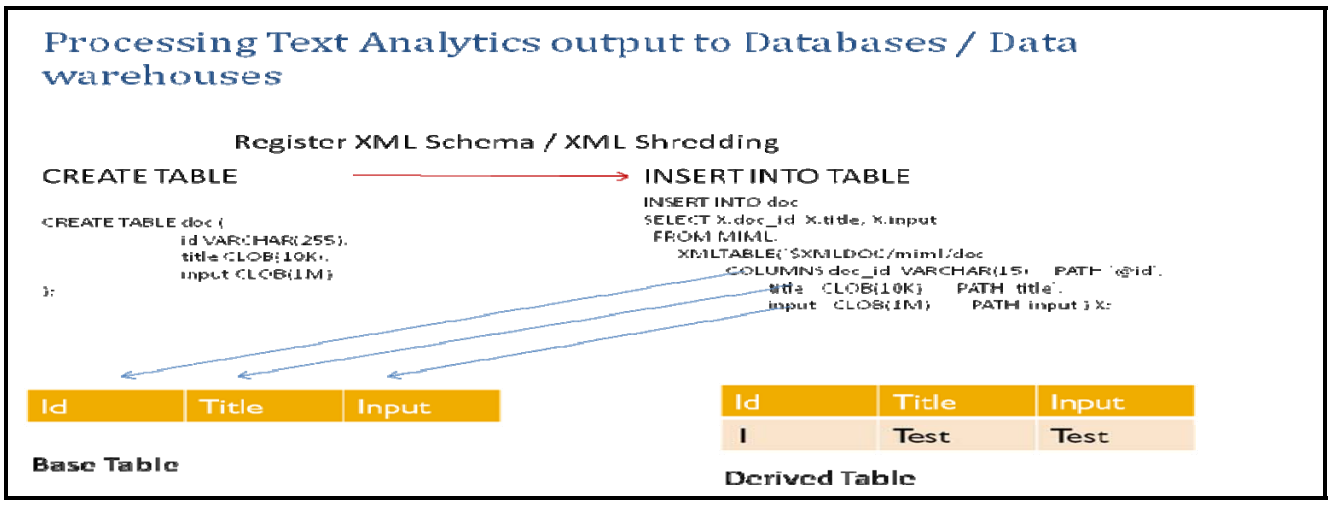


Figure V - XML Storage and process

XI. LIMITATIONS

1. Extracting textual sources from different sources requires lot of external software programming to be involved based on source.
2. Management in master data for unstructural data coming from various sources.

XII CONCLUSIONS

Text Analytics is playing a major role in enabling support to be processed in decision support or transaction based systems by removing the barrier between structured and unstructured data significantly impacts the way companies treat and govern data. By its nature, unstructured data makes it difficult to directly extract meaningful information and combine it with structured data sources. How ever, structure and semantic information can be added to unstructured data through text tagging and annotation, making it suitable for integration with other data sources..

References:

[1] Fan, W., Wallace,L., Rich,S., and Zhang, Z., 2006 Tapping the power of text mining, proceeding of communications of the ACM, pp.78-82.
[2] Grimes,S., 2008 Text Analytics for Dummies, proceedings from text analytics summit 2008 workshop on text analytics, pp11-16.
[3] Nicola,M., Summerland M., and Zeidenstien, k., 2008 From Text Analytics to Data Warehouses, IBM Db2 developer works, pp-124.
[4] BI search and Text Analytics.
[5] http://www.tdwi.org/Publications/WhatWorks/display.aspx?id=8449, (2007)
[6] Hearst, M. What is text mining. http://www.sims.berkeley.edu/~hearst/textmining.html, (2004)
[7] Kuo, J., 2007 Turning Unstructured text into insight, Business Objects White paper, pp. 4-8
[8] Surekha, Asish., Sreekumar, Sukumaran., 2005 Integrating Structured and Unstructured Data Using Text tagging and annotation,

Appendix

[DB2 Support based queries]
SQL (Structured Query Language) definitions for creating tables and Inserting rows in to base tables

```
CREATE TABLE CATEGORY_SUBJECTDESC' (
   'ID' INTEGER NOT NULL,
   'KEYWORD' VARCHAR(255) NOT NULL
         )
    ENGINE = INNODB;

CREATE TABLE DATE (
   ID INTEGER NOT NULL,
   DATE DATE NOT NULL,
   YEAR SMALLINT NOT NULL,
   QUARTER SMALLINT NOT NULL,
   MONTH SMALLINT NOT NULL,
   DAY_OF_YEAR SMALLINT NOT NULL,
   DAY_OF_MONTH SMALLINT NOT NULL,
   DAY_OF_WEEK SMALLINT NOT NULL,
   WEEK_OF_YEAR SMALLINT NOT NULL,
   WEEK_OF_MONTH SMALLINT NOT NULL
      )
      ;

CREATE TABLE DOCUMENT (
   ID INTEGER NOT NULL,
   DOC_ID VARCHAR(255) NOT NULL,
   TITLE CLOB(30720) NOT NULL
   NOT LOGGED,
   INPUT CLOB(1048576) NOT NULL
   NOT LOGGED
      )
      ;

CREATE TABLE DOCUMENT_ANALYSIS (
   DOCUMENT INTEGER NOT NULL,
   DATE INTEGER NOT NULL,
   UCAT INTEGER NOT NULL,
   CATEGORY_SUBJECTDESC INTEGER
   DEFAULT 0 NOT NULL
      )
      ;
```

```
INSERT INTO doc
SELECT X.doc_id, X.title, X.input
  FROM MIML,
   XMLTABLE('$XMLDOC/miml/doc'
    COLUMNS doc_id  VARCHAR(15)
    PATH '@id',
    title   CLOB(10K)   PATH 'title',
    input   CLOB(1M) PATH 'input') X;

INSERT INTO standard_kw
SELECT X.doc_id, X.cat, X.val
  FROM MIML,

XMLTABLE('$XMLDOC/miml/doc/feature
[@type="standard"]/kw'
COLUMNS doc_id  VARCHAR(15)
PATH '../../@id',
cat     VARCHAR(50)   PATH '@cat',
val     VARCHAR(4096) PATH '@val') x,
```

Kalli Srinivasa Nageswara Prasad has completed M.Sc(Tech)., M.Sc., M.S (Software Systems)., P.G.D.C.S. He is currently pursuing Ph.D degree in the field of Data Mining at Sri Venkateswara University, Tirupathi, Andhra Pradesh State, India.

S.Ramakrishna is currently working as a professor in the Department of Computer Science, College of Commerce, Management & Computer Sciences in Sri Venkateswara university, Tirupathi, Andhra Pradesh State, India. He has completed M.Sc, M.Phil., Ph.D., M.Tech(IT).He is specialized in Fluid Dynamics & Theoretical Computer Science. His area of Research includes Artificial Intelligence , Data Mining & Networks. He has an experience of 23 years in Teaching field. He has published 43 Research papers in National & International Journals. He has also attended 13 national Conferences and 11 International Conferences. He has guided 12 Ph.D Scholars and 15 M.Phil Scholars