A Survey of Text Mining Techniques and Applications

Vishal Gupta
Lecturer Computer Science & Engineering, University Institute of Engineering & Technology,
Panjab University Chandigarh, India
Email: vishal@pu.ac.in

Gurpreet S. Lehal
Professor & Head, Department of Computer Science, Punjabi University Patiala, India
Email: gslehal@yahoo.com

***Abstract*— Text Mining has become an important research area. Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. In this paper, a Survey of Text Mining techniques and applications have been s presented.**

***Index Terms*—text mining, information extraction, topic tracking, summarization, clustering, question answering etc.**

I. Introduction

Text Mining [1] is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down.

Text mining is a variation on a field called data mining [2],that tries to find interesting patterns from large databases. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value. Knowledge may be discovered from many sources of information, yet, unstructured texts remain the largest readily available source of knowledge.

The problem of Knowledge Discovery from Text (KDT) [6] is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process. KDT plays an increasingly significant role in emerging applications, such as Text Understanding.

Text mining [1] is similar to data mining, except that data mining tools [2] are designed to handle structured data from databases , but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. As a result, text mining is a much better solution for companies. To date, however, most research and development efforts have centered on data mining efforts using structured data. The problem introduced by text mining is obvious: natural language was developed for humans to communicate with one another and to record information, and computers are a long way from comprehending natural language.

Humans have the ability to distinguish and apply linguistic patterns to text and humans can easily overcome obstacles that computers cannot easily handle such as slang, spelling variations and contextual meaning. However, although our language capabilities allow us to comprehend unstructured data, we lack the computer's ability to process text in large volumes or at high speeds. Figure 1 on next page, depicts a generic process model [3] for a text mining application.

Starting with a collection of documents, a text mining tool would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. Three text analysis techniques are shown in the example, but many other combinations of techniques could be used depending on the goals of the organization. The resulting information can be placed in a management information

Manuscript received March 26, 2009; revised May 15, 2009; accepted June 10, 2009.
Corresponding author: Vishal Gupta

system, yielding an abundant amount of knowledge for the user of that system.

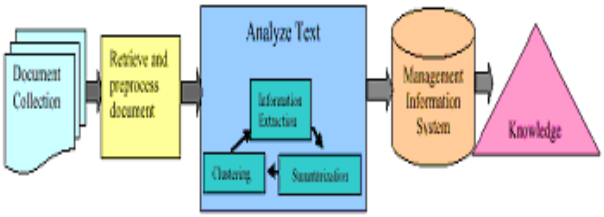

Figure 1. An example of Text Mining

II. TECHNOLOGY FOUNDATIONS

Although the differences in human and computer languages are expansive, there have been technological advances which have begun to close the gap. The field of natural language processing has produced technologies that teach computers natural language so that they may analyze, understand, and even generate text. Some of the technologies [3] that have been developed and can be used in the text mining process are information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization, and question answering. In the following sections we will discuss each of these technologies and the role that they play in text mining. We will also illustrate the type of situations where each technology may be useful in order to help readers identify tools of interest to themselves or their organizations.

A. Information Extraction

A starting point for computers to analyze unstructured text is to use information extraction. Information extraction software identifies key phrases and relationships within text. It does this by looking for predefined sequences in text, a process called pattern matching. The software infers the relationships between all the identified people, places, and time to provide the user with meaningful information. This technology can be very useful when dealing with large volumes of text. Traditional data mining assumes that the information to be "mined" is already in the form of a relational database. Unfortunately, for many applications, electronic information is only available in the form of free natural-language documents rather than structured databases. Since IE addresses the problem of transforming a corpus of textual documents into a more structured database, the database constructed by an IE module can be provided to the KDD module for further mining of knowledge as illustrated in Figure 2.

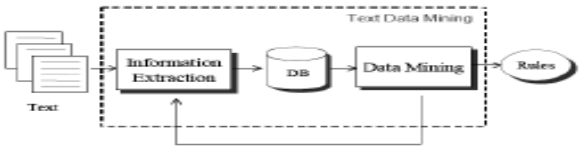

Figure 2. Overview of IE-based text mining framework

After mining knowledge from extracted data, DISCOTEX [11] can predict information missed by the previous extraction using discovered rules.

First, we show the pseudo code [11] for the rule mining phase in Figure 3. A final step shown in the figure is filtering the discovered rules on both the training data and a disjoint set of labeled validation data in order to retain only the most accurate of the induced rules. IE that make any incorrect predictions on either the training or validation extracted templates are discarded. Since association rules are not intended to be used together as a set as classification rules are, we focus on mining prediction rules for this task.

Input: $\mathcal{D}$ is the set of document.
Output: RB is the set of prediction rules.
Function RuleMining ($\mathcal{D}$)
Determine T, a threshold value for rule validation
Create a database of labeled examples (by applying IE to the document corpus, $\mathcal{D}$)
For each labeled example $D \in \mathcal{D}$ **do**
  F := set of slot fillers of D
  Convert F to binary features
Build a prediction rule base, RB (by applying rule miner to the binary data, F)
For each prediction rule $R \in RB$ **do**
  Verify R on training data and validation data
  If the accuracy of R is lower than T
    Delete R from RB
Return RB.

Figure 3. Algorithm specification: rule mining

The extraction algorithm [11] which attempts to improve recall by using the mined rules is summarized in figure 4. Note that the final decision whether or not to extract a predicted filler is based on whether the filler (or any of its synonyms) occurs in the document as a substring. If the filler is found in the text, the extractor considers its prediction confirmed and extracts the filler.

Input: RB is the set of prediction rules.
  $\mathcal{D}$ is the set of documents.
Output: F is the set of slot fillers extracted.
Function InformationExtraction (RB, $\mathcal{D}$)
$F := \emptyset$.
For each example $D \in \mathcal{D}$ **do**
  Extract fillers from D using extraction rules and add them to F
  For each rule R in the prediction rule base RB **do**
    If R fires on the current extracted fillers
      If the predicted filler is a substring of D
        Extract the predicted filler and add it to F
Return F.

Figure 4. Algorithm specification

B. Topic Tracking

A topic tracking system works by keeping user profiles and, based on the documents the user views, predicts other documents of interest to the user. Yahoo offers a free topic tracking tool (www.alerts.yahoo.com) that allows users to choose keywords and notifies them when news relating to those topics becomes available. Topic tracking technology does have limitations, however. For example, if a user sets up an alert for "text mining", s/he will receive several news stories on mining for minerals, and very few that are actually on text

mining. Some of the better text mining tools let users select particular categories of interest or the software automatically can even infer the user's interests based on his/her reading history and click-through information.

There are many areas where topic tracking can be applied in industry. It can be used to alert companies anytime a competitor is in the news. This allows them to keep up with competitive products or changes in the market. Similarly, businesses might want to track news on their own company and products. It could also be used in the medical industry by doctors and other people looking for new treatments for illnesses and who wish to keep up on the latest advancements. Individuals in the field of education could also use topic tracking to be sure they have the latest references for research in their area of interest.

Keywords are a set of significant words in an article that gives high-level description of its contents to readers. Identifying keywords from a large amount of on-line news data is very useful in that it can produce a short summary of news articles. As on-line text documents rapidly increase in size with the growth of WWW, keyword extraction [13] has become a basis of several text mining applications such as search engine, text categorization, summarization, and topic detection. Manual keyword extraction is an extremely difficult and time consuming task; in fact, it is almost impossible to extract keywords manually in case of news articles published in a single day due to their volume. For a rapid use of keywords, we need to establish an automated process that extracts keywords from news articles. The architecture of keyword extraction system is presented in figure 5. HTML news pages are gathered from a Internet portal site. And candidate keywords are extracted throw keyword extraction module. And finally keywords are extracted by cross-domain comparison module. Keyword extraction module is described in detail. We make tables for 'document', 'dictionary', 'term occur fact' and 'TF-IDF weight' in relational database. At first the downloaded news documents are stored in 'Document' table and nouns are extracted from the documents in 'Document table.

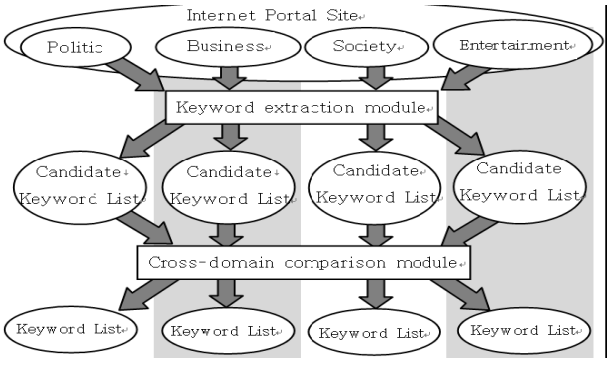

Figure 5. The architecture of keyword extraction system

Then the facts which words are appeared in documents are updated to 'Term occur fact' table. Next, TF-IDF weights for each word are calculated using 'Term occur fact' table and the result are updated to 'TF-IDF weight' table. Finally, using 'TF-IDF weight' table, 'Candidate keyword list' for each news domain with words is ranked high. Keyword extraction module is given in figure 6.

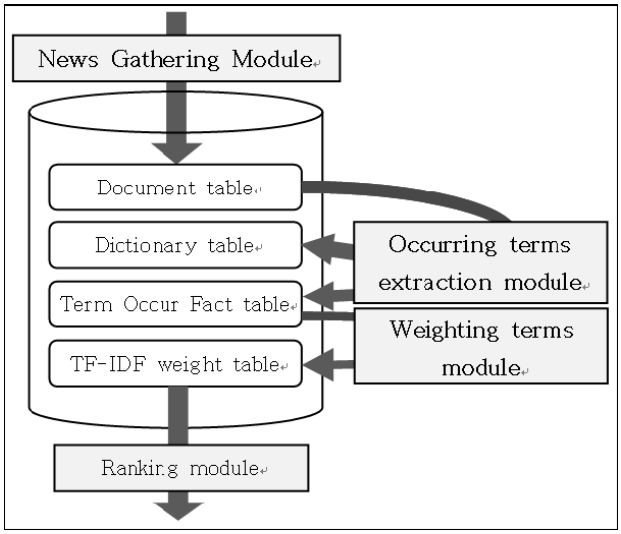

Figure 6. Keyword extraction module

Lexical chaining [14] is a method of grouping lexically related terms into so called lexical chains. Topic tracking involves tracking a given news event in a stream of news stories i.e. finding all the subsequent stories in the news stream.

In multi vector [15] topic tracking system proper names, locations and normal terms are extracted into distinct sub vectors of document representation. Measuring the similarity of two documents is conducted by comparing two sub-vectors at a time. Number of features that effect the performance of topic tracking system are analyzed. First choice is to choose one characteristic, such as the choice of words, words or phrases such as string as a feature in this term to make features as an example. that discuss the given event.

C. Summarization

Text summarization is immensely helpful for trying to figure out whether or not a lengthy document meets the user's needs and is worth reading for further information. With large texts, text summarization software processes and summarizes the document in the time it would take the user to read the first paragraph. The key to summarization is to reduce the length and detail of a document while retaining its main points and overall meaning. The challenge is that, although computers are able to identify people, places, and time, it is still difficult to teach software to analyze semantics and to interpret meaning.

Generally, when humans summarize text, we read the entire selection to develop a full understanding, and then write a summary highlighting its main points. Since computers do not yet have the language capabilities of humans, alternative methods must be considered. One of the strategies most widely used by text summarization tools, sentence extraction, extracts important sentences

from an article by statistically weighting the sentences. Further heuristics such as position information are also used for summarization.

For example, summarization tools may extract the sentences which follow the key phrase "in conclusion", after which typically lie the main points of the document. Summarization tools may also search for headings and other markers of subtopics in order to identify the key points of a document. Microsoft Word's AutoSummarize function is a simple example of text summarization. Many text summarization tools allow the user to choose the percentage of the total text they want extracted as a summary. Summarization can work with topic tracking tools or categorization tools in order to summarize the documents that are retrieved on a particular topic. If organizations, medical personnel, or other researchers were given hundreds of documents that addressed their topic of interest, then summarization tools could be used to reduce the time spent sorting through the material. Individuals would be able to more quickly assess the relevance of the information to the topic they are interested in.

An automatic summarization [16] process can be divided into three steps: (1)In the preprocessing step a structured representation of the original text is obtained; (2) In the processing step an algorithm must transform the text structure into a summary structure; and (3) In the generation step the final summary is obtained from the summary structure. The methods of summarization can be classified, in terms of the level in the linguistic space, in two broad groups: (a) shallow approaches, which are restricted to the syntactic level of representation and try to extract salient parts of the text in a convenient way; and (b) deeper approaches, which assume a semantics level of representation of the original text and involve linguistic processing at some level.

In the first approach the aim of the preprocessing step is to reduce the dimensionality of the representation space, and it normally includes: (i) stop-word elimination –common words with no semantics and which do not aggregate relevant information to the task (e.g., "the", "a") are eliminated; (ii) case folding: consists of converting all the characters to the same kind of letter case - either upper case or lower case; (iii) stemming: syntactically-similar words, such as plurals, verbal variations, etc. are considered similar; the purpose of this procedure is to obtain the stem or radix of each word, which emphasize its semantics. A frequently employed text model is the vector model. After the preprocessing step each text element –a sentence in the case of text summarization – is considered as a N-dimensional vector. So it is possible to use some metric in this space to measure similarity between text elements. The most employed metric is the cosine measure, defined as cos q = (<x.y>) / (|x| . |y|) for vectors x and y, where (<,>) indicates the scalar product, and |x| indicates the module of x. Therefore maximum similarity corresponds to cos q = 1, whereas cos q = 0 indicates total discrepancy between the text elements. To implement text summarization based on fuzzy logic, MATLAB is usually used since it is possible to simulate fuzzy logic in this software. Select characteristic of a text such as sentence length, similarity to little, similarity to key word and etc. as the input of fuzzy system. Then, all the rules needed for summarization are entered in the knowledge base of this system. After ward, a value from zero to one is obtained for each sentence in the output based on sentence characteristics and the available rules in the knowledge base. The obtained value in the output determines the degree of the importance of the sentence in the final summary.

The Kernel of generating text summary using sentence-selection based text summarization approach [17] is shown in figure7.

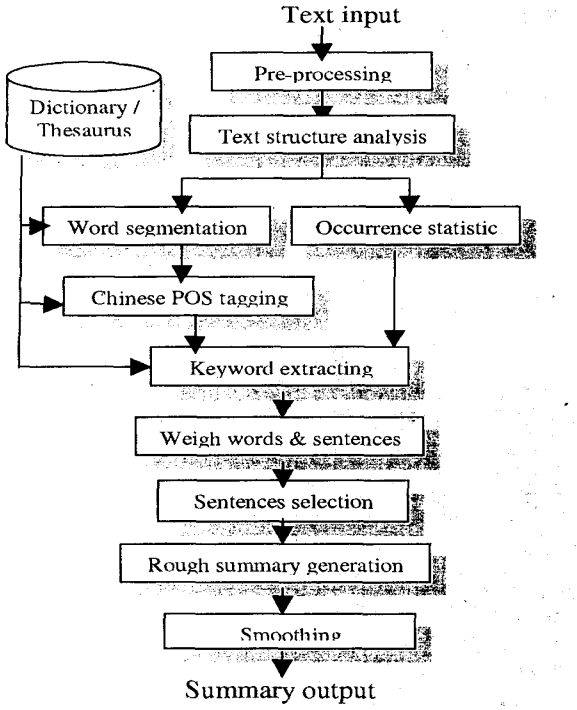

Figure 7. Kernel of text summarization

D. Categorization

Categorization involves identifying the main themes of a document by placing the document into a pre-defined set of topics. When categorizing a document, a computer program will often treat the document as a "bag of words." It does not attempt to process the actual information as information extraction does. Rather, categorization only counts words that appear and, from the counts, identifies the main topics that the document covers. Categorization often relies on a thesaurus for which topics are predefined, and relationships are identified by looking for broad terms, narrower terms, synonyms, and related terms. Categorization tools normally have a method for ranking the documents in order of which documents have the most content on a particular topic.

As with summarization, categorization can be used with topic tracking to further specify the relevance of a document to a person seeking information on a topic. The documents returned from topic tracking could be ranked

by content weights so that individuals could give priority to the most relevant documents first. Categorization can be used in a number of application domains. Many businesses and industries provide customer support or have to answer questions on a variety of topics from their customers. If they can use categorization schemes to classify the documents by topic, then customers or end-users will be able to access the information they seek much more readily. The goal of text categorization is to classify a set of documents into a fixed number of predefined categories. Each document may belong to more than one class.

Using supervised learning algorithms [18], the objective is to learn classifiers from known examples (labeled documents) and perform the classification automatically on unknown examples (unlabeled documents). Figure.8 shows the overall flow diagram of the text categorization task. Consider a set of labeled documents from a source $D = [d1,d2,....dn]$ belonging to a set of classes $C = [c1,c2,...,cp]$. The text categorization task is to train the classifier using these documents, and assign categories to new documents. In the training phase, the *n* documents are arranged in *p* separate folders, where each folder corresponds to one class. In the next step, the training data set is prepared via a feature selection process.

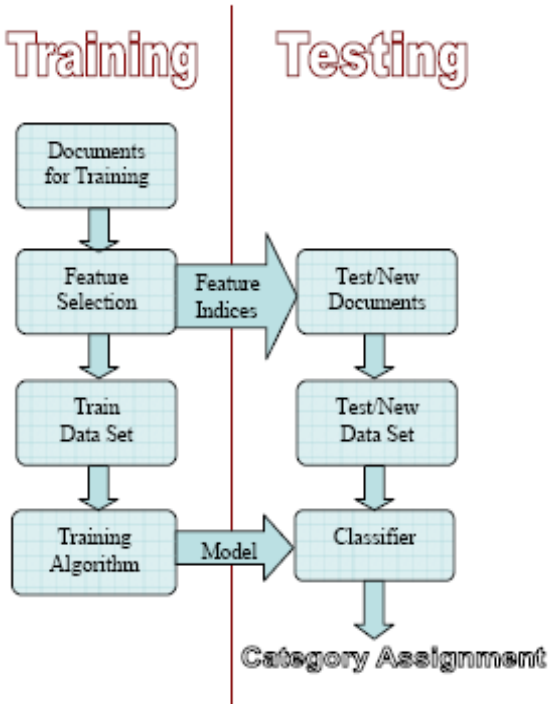

Figure 8. Flow Diagram of Text Categorization

Text data typically consists of strings of characters, which are transformed into a representation suitable for learning. It is observed from previous research that words work well as features for many text categorization tasks. In the feature space representation, the sequences of characters of text documents are represented as sequence of words. Feature selection involves tokenizing the text, indexing and feature space reduction. Text can be tokenized using term frequency (TF), inverse document frequency (IDF), term frequency inverse document frequency (TFIDF) or using binary representation. Using these representations the global feature space is determined from entire training document collection.

In text manifold [19] categorization method, the text documents are treated as vectors in an n-dimensional space, where every dimension corresponds to a term. Then the metrics such as the cosine of the angle between two documents can be defined. However this space may be intrinsically located on the low dimensional manifold. The metric therefore should be defined according to the properties of manifold so as to improve the text categorization furthermore. The whole process is illustrated as Figure 9.

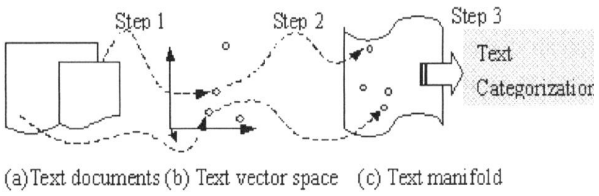

Figure 9. Framework of text categorization on text manifold

In TCBPLK [20] method of text categorization, texts are automatically assigned to appointed species according to text content. Similar texts are assigned to the same species through calculating the similarity among the texts. After the process of pattern aggregation for the word matrix, the numbers of words are greatly decreased, then TF.IDF method is applied to constructing the VSM. As the dimensions of the text are greatly decreased through the process of the P-L, the method decreases the learning time, and advances the speed and the of text categorization.

E. Clustering

Clustering [7] is a technique used to group similar documents, but it differs from categorization in that documents are clustered on the fly instead of through the use of predefined topics. Another benefit of clustering is that documents can appear in multiple subtopics, thus ensuring that a useful document will not be omitted from search results. A basic clustering algorithm creates a vector of topics for each document and measures the weights of how well the document fits into each cluster. Clustering technology can be useful in the organization of management information systems, which may contain thousands of documents.

In K-means clustering algorithm[21] , while calculating Similarity between text documents, not only consider eigenvector based on algorithm of term frequency statistics ,but also combine the degree of association between words ,then the relationship between keywords has been taken into consideration ,thereby it lessens sensitivity of input sequence and frequency, to a certain extent, it considered semantic understanding , effectively raises similarity accuracy of small text and simple sentence as well as preciseness and recall rate of

text cluster result .The algorithm model with the idea of co-mining shows as Fig 10.

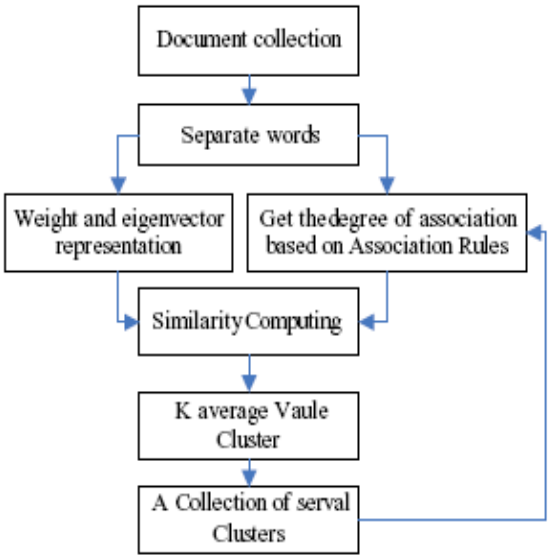

Figure 10. The flow diagram of K-means clustering based on co-mining

In word relativity-based clustering (WRBC) method [22], text clustering process contains four main parts: text reprocessing, word relativity computation, word clustering and text classification. See Figure 11.

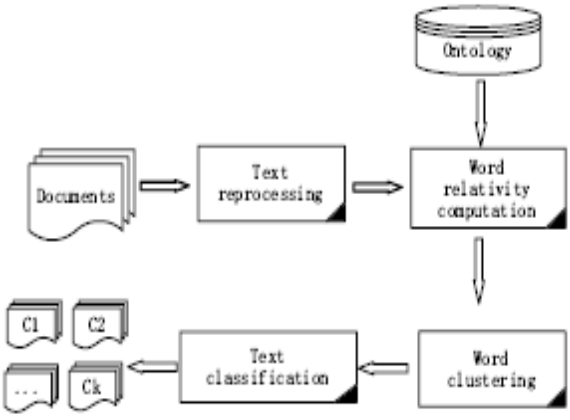

Figure 11. Word relativity-based clustering method

The first step in text clustering is to transform documents, which typically are strings of characters into a suitable representation for the clustering task.

(1)Remove stop-words: The stop-words are high frequent words that carry no information (i.e. pronouns, prepositions, conjunctions etc.). Remove stop-words can improve clustering results.

(2) Stemming: By word stemming it means the process of suffix removal to generate word stems. This is done to group words that have the same conceptual meaning, such as work, worker, worked and working.

(3) Filtering: Domain vocabulary V in ontology is used for filtering. By filtering, document is considered with related domain words (term). It can reduce the documents dimensions. A central problem in statistical text clustering is the high dimensionality of the feature space. Standard clustering techniques cannot deal with such a large feature set, since processing is extremely costly in computational terms. We can represent documents with some domain vocabulary in order to solving the high dimensionality problem. In the beginning of word clustering, one word randomly is chosen to form initial cluster. The other words are added to this cluster or new cluster, until all words are belong to m clusters. This method allow one word belong to many clusters and accord with the fact. This method implements word clustering by calculating word relativity and then implements text classification.

F. Concept Linkage

Concept linkage tools [3] connect related documents by identifying their commonly-shared concepts and help users find information that they perhaps wouldn't have found using traditional searching methods. It promotes browsing for information rather than searching for it. Concept linkage is a valuable concept in text mining, especially in the biomedical fields where so much research has been done that it is impossible for researchers to read all the material and make associations to other research. Ideally, concept linking software can identify links between diseases and treatments when humans can not. For example, a text mining software solution may easily identify a link between topics X and Y, and Y and Z, which are well-known relations. But the text mining tool could also detect a potential link between X and Z, something that a human researcher has not come across yet because of the large volume of information s/he would have to sort through to make the connection.

G. Information Visualization

Visual text mining, or information visualization [3], puts large textual sources in a visual hierarchy or map and provides browsing capabilities, in addition to simple searching. DocMiner as shown in figure12, is a tool that shows mappings of large amounts of text, allowing the user to visually analyze the content. The user can interact with the document map by zooming, scaling, and creating sub-maps. Information visualization is useful when a user needs to narrow down a broad range of documents and explore related topics. The government can use information visualization to identify terrorist networks or to find information about crimes that may have been previously thought unconnected. It could provide them, with a map of possible relationships between suspicious activities so that they can investigate connections that they would not have come up with on their own.

The goal of information visualization, the construction may be conducted into three steps: (1) Data preparation: i.e. determine and acquire original data of visualization and form original data space. (2) Data analysis and extraction: i.e. analyze and extract visualization data needed from original data and form visualization data space. (3) Visualization mapping: i.e. employ certain mapping algorithm to map visualization data space to

visualization target. InforVisModel [23] divide the construction into five steps:

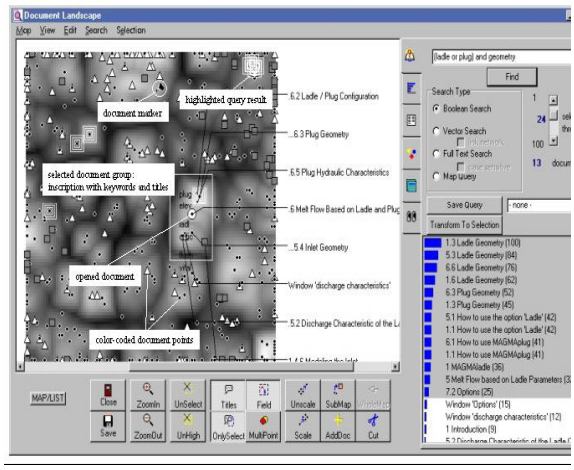

Figure 12. Doc Miner's interface

(1) Information collection: to collect information resources needed from databases or WWW.

(2) Information indexing: to index collected information resources to form original data sources.

(3) Information retrieval: to query information lists in conformity to result from original data sources according to the need of retrieval.

(4) Generation of visualization data: to transform data in the retrieved results into visualization data.

(5) Display of visualization interface: to map visualization data to visualization target and display them on visualization interface. InfoVisModel visualization model is shown in figure13.

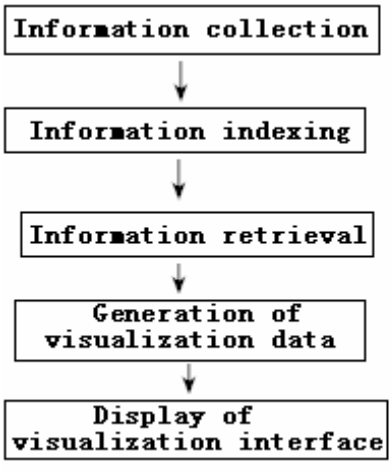

Figure 13. InfoVisModel visualization model

H. Question Answering

Another application area of natural language processing is natural language queries, or question answering (Q&A), which deals with how to find the best answer to a given question. Many websites that are equipped with question answering technology, allow end users to "ask" the computer a question and be given an answer. Q&A can utilize multiple text mining techniques. For example, it can use information extraction to extract entities such as people, places, events; or question categorization to assign questions into known types (who, where, when, how, etc.). In addition to web applications, companies can use Q&A techniques internally for employees who are searching for answers to common questions. The education and medical areas may also find uses for Q&A in areas where there are frequently asked questions that people wish to search.

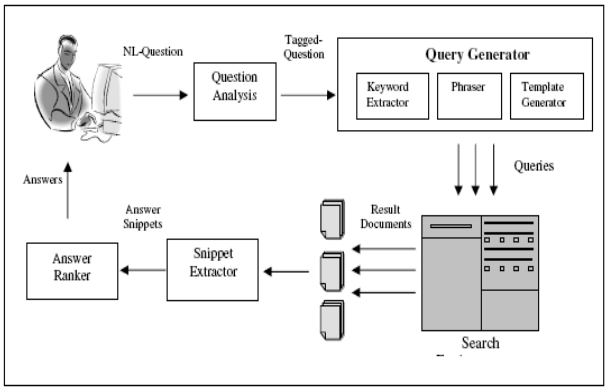

Figure 14. Architecture of Question answering system

Figure14 shows the architecture of question answering system.[24] The system takes in a natural language (NL) question in English from the user. This question is then passed to a Part-of-Speech (POS) tagger which parses the question and identifies POS of every word involved in the question. This tagged question is then used by the query generators which generate different types of queries, which can be passed to a search engine. These queries are then executed by a search engine in parallel. The search engine provides the documents which are likely to have the answers we are looking for. These documents are checked for this by the answer extractor. Snippet Extractor extracts snippets which contain the query phrases/words from the documents. These snippets are passed to the ranker which sorts them according to the ranking algorithm.

In QUASAR [25] System, user provides a question and this is handed over to the Question Analysis and Passage Retrieval modules. Next, the Answer Extraction obtains the answer from the expected type, constraints and passages returned by Question Analysis and *Passage Retrieval* modules. The architecture is shown in figure15. The main objective of question analysis module is to derive the expected answer type from the question text. This is a crucial step of the processing since the Answer Extraction module uses a different strategy depending on the expected answer type. Another operation performed by this module is to analyze the query with the purpose of identifying the constraints to be used in the AE phase. These constraints are made by sequences of words extracted from the POS-tagged query by means of POS patterns and rules. For instance, any sequence of nouns (such as "ozone hole") is considered as a relevant pattern.

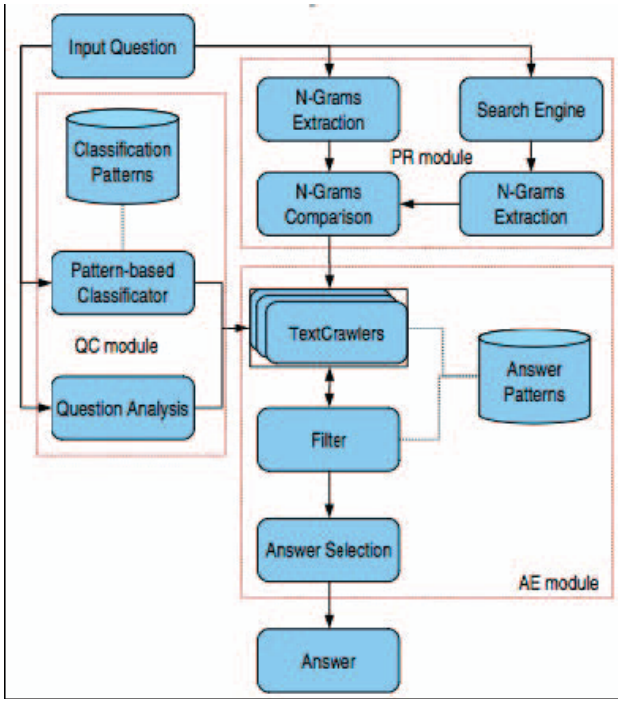

Figure 15. QUASAR QUESTION ANSWERING SYSTEM

In Passage Retrieval module, user question is handed over also to the JIRS Passage Retrieval system, more specifically to its Search Engine and N-grams Extraction components. Passages with the relevant terms (i.e., without stop words) are found by the Search Engine using the classical IR system. Sets of 1-grams, 2-grams,. . ., *n*-grams are extracted from the extended passages and from the user question. In both cases, *n* will be the number of question terms. A comparison between the *n*-gram sets from the passages and the user question is done in order to obtain the weight of each passage. The weight of a passage will be heavier if the passage contains greater *n*-gram structures of the question.

The input of answer extraction module is constituted by the *n* passages returned by the PR module and the constraints (including the expected type of the answer) obtained through the Question Analysis module.

A Text Crawler is instantiated for each of the *n* passages with a set of patterns for the expected type of the answer and a pre-processed version of the passage text. The pre-processing of passage text consists in separating all the punctuation characters from the words and in stripping off the annotations of the passage. It is important to keep the punctuation symbols because we observed that they usually offer important clues for the individuation of the answer: The Text Crawler begins its work by searching all the passage's substrings matching the expected answer pattern. Then a weight is assigned to each found substring *s*, depending on the positions of the constraints, if *s* does not include any of the constraint words. The filter module takes advantage of some knowledge resources, such as a mini knowledge base or the web, in order to discard the candidate answers which do not match with an allowed pattern or that do match with a forbidden pattern. For instance, a list of country names in the four languages has been included in the knowledge base in order to filter country names when looking for countries. When the filter rejects a candidate, the Text Crawler provide it with the next best weighted candidate, if there is one. Finally, when all Text Crawlers end their analysis of the text, the Answer Selection module selects the answer to be returned by the system.

I. Association Rule Mining

Association rule mining (ARM) [33] is a technique used to discover relationships among a large set of variables in a data set. It has been applied to a variety of industry settings and disciplines but has, to date, not been widely used in the social sciences, especially in education, counseling, and associated disciplines. ARM refers to the discovery of relationships among a large set of variables, that is, given a database of records, each containing two or more variables and their respective values, ARM determines variable-value combinations that frequently occur. Similar to the idea of correlation analysis (although they are theoretically different), in which relationships between two variables are uncovered, ARM is also used to discover variable relationships, but each relationship (also known as an association rule) may contain two or more variables.

ARM has been extensively employed in business decision- making processes. ARM discovers what items customers typically purchase together. These associations can then be used by a supermarket, for example, to place frequently co-purchased items in adjacent shelves to increase sales. Thus, if bread and cereal are often purchased together, placing these items in close proximity may encourage customers to buy them within single visits to the supermarket. ARM is a technique that is part of the field of data mining. Also known as knowledge discovery in databases.

In Association Rules for Text Mining [34], The focus is to study the relationships and implications among *topics*, or descriptive concepts, that are used to characterize a corpus. The goal is to discover important association rules within a corpus such that the presence of a set of topics in an article implies the presence of another topic. For example, one might learn in headline news that whenever the words "Greenspan" and "inflation" occur, it is highly probably that the stock market is also mentioned. Figure15a shows a high-level system overview of the topic association mining system. A corpus of narrative text is fed into a text engine for topic extractions. The mining engine then reads the topics from the text engine and generates topic association rules. Finally, the resultant association rules are sent to the visualization system for further analysis.

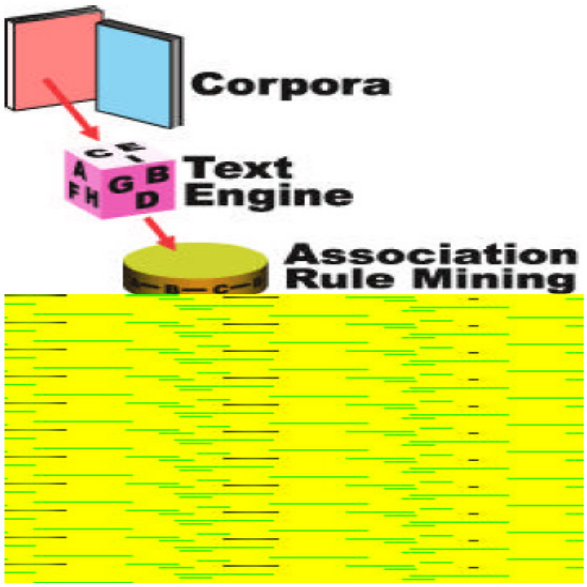

Figure 15a. Overview of our topic association system

III. TEXT MINING APPLICATIONS

The main Text Mining applications [4] are most often used in the following sectors:

- Publishing and media.
- Telecommunications, energy and other services industries.
- Information technology sector and Internet.
- Banks, insurance and financial markets.
- Political institutions, political analysts, public administration and legal documents.
- Pharmaceutical and research companies and healthcare.

The sectors analyzed are characterized by a fair variety in the applications being experimented. however, it is possible to identify some sectorial specifications in the use of TM, linked to the type of production and the objectives of the knowledge management leading them to use TM. The publishing sector, for example, is marked by prevalence of Extraction Transformation Loading applications for the cataloguing, producing and the optimization of the information retrieval.

In the banking and insurance sectors, on the other hand, CRM applications are prevalent and aimed at improving the management of customer communication, by automatic systems of message re-routing and with applications supporting the search engines asking questions in natural language. In the medical and pharmaceutical sectors, applications of Competitive Intelligence and Technology Watch are widespread for the analysis, classification and extraction of information from articles, scientific abstracts and patents. A sector in which several types of applications are widely used is that of the telecommunications and service companies: the most important objectives of these industries are that all applications find an answer, from market analysis to human resources management, from spelling correction to customer opinion survey.

A. Text Mining Applications in Knowledge and Human Resource management

Text Mining is widely used in field of knowledge and Human Resource Management. Following are its few applications in these areas:

1) Competitive Intelligence: The need to organize and modify their strategies according to demands and to the opportunities that the market present requires that companies collect information about themselves, the market and their competitors, and to manage enormous amount of data, and analyzing them to make plans. The aim of Competitive Intelligence [4] is to select only relevant information by automatic reading of this data. Once the material has been collected, it is classified into categories to develop a database, and analyzing the database to get answers to specific and crucial information for company strategies.

The typical queries concern the products, the sectors of investment of the competitors, the partnerships existing in markets, the relevant financial indicators, and the names of the employees of a company with a certain profile of competences. Before the introduction of TM, there was a division that was entirely dedicated to the continuous monitoring of information (financial, geopolitical, technical and economic) and answering the queries coming from other sectors of the company. In these cases the return on investment by the use of TM technologies was self evident when compared to results previously achieved by manual operators.

In some cases, if a scheme of categories is not defined a priori, cauterization procedures are used to classify the set of documents (considered) relevant with regard to a certain topic, in clusters of documents with similar contents. The analysis of the key concepts present in the single clusters gives an overall vision of the subjects dealt with in the single texts.

More company and news information are increasingly available on the web. As such, it has become a gold mine of online information that is crucial for competitive intelligence (CI). To harness this information, various search engines and text mining techniques have been developed to gather and organize it. However, the user has no control on how the information is organized through these tools and the information clusters generated may not match their needs. The process of manually compiling documents according to a user's needs and preferences and into actionable reports is very labour intensive, and is greatly amplified when it needs to be updated frequently. Updates to what has been collected often require a repeated search, filtering of previously retrieved documents and re-organizing.

FOCI [26] (Flexible Organizer for Competitive Intelligence), can help the knowledge worker in the gathering, organizing, tracking, and dissemination of competitive intelligence or knowledge bases on the web. FOCI allows a user to define and personalize the organization of the information clusters according to their needs and preferences into portfolios. Figure16 shows the architecture of FOCI. It comprises an Information

Gathering module for retrieving relevant information from the web sources; a Content Management module for organizing information into portfolios and personalizing the portfolios; a Content Mining module for discovering new information and a Content Publishing module for publishing and sharing of information and a user interface front end for graphical visualization and users interactions. The portfolios created are stored into CI knowledge bases which can be shared by the users within an organization.

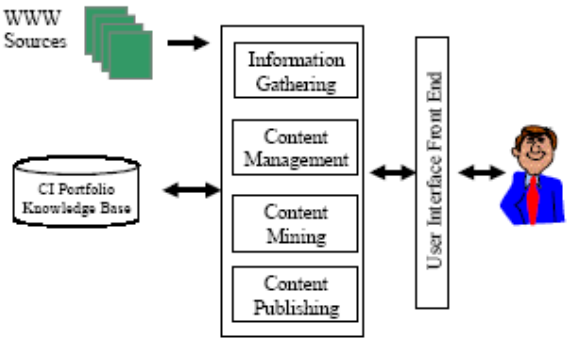

Figure.16. FOCI System Architecture

Text mining can represent flexible approaches to information management, research and analysis. Thus text mining can expand the fists of data mining to the ability to deal with textual materials. The following Fig. 17 addresses the process of using text mining and related methods and techniques to extract business intelligence [27] from multi sources of raw text information. Although there seems something like that of data mining, this process of text mining gains the extra power to extract expanding business intelligence.

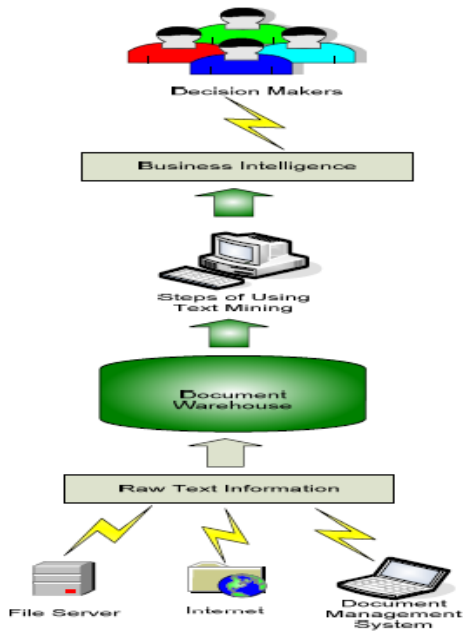

Figure 17. Text Mining in Business Intelligence

2) Extraction Transformation Loading: Extraction Transformation Loading [4] are aimed at filing non-structured textual material into categories and structured fields. The search engines are usually associated with ETL that guarantee the retrieval of information, generally by systems foreseeing conceptual browsing and questioning in natural language. The applications are found in the editorial sector, the juridical and political document field and medical health care. In the legal documents sector the document filing and information management operations deal with the particular features of language, in which the identification and tagging of relevant elements for juridical purposes is necessary.

The data can come from any source i.e., a mainframe application, an ERP application, a CRM tool, a flat file, and an Excel spreadsheet—even a message queue. All these types of data must be transformed into a single suitable format and stored in large repository called Data warehouse. To make a Data warehouse we have to follow a process known as Extraction, transformation, and loading (ETL) [28] which involves

- Extracting data from various outside sources.
- Transforming it to fit business needs, and ultimately
- Loading it into the data warehouse.

The first part of an ETL process is to extract the data from various source systems. Data warehouse consolidate data from different source systems. These sources may have different formats of data. Data source formats can be relational databases and flat files, non-relational database structures such as IMS or other data structures such as VSAM or ISAM So Extraction of these different format data which uses different internal representation is difficult process. Extraction tool must understand all different data storage formats.

The transformation phase applies a number of rules to the extracted data so as to convert different data formats into single format. These transformation rules will be applied by transformation tool as per the requirements. Following transformations types may be required:

- Selecting only those which don't have null values.
- Translating coded values
- Making all same values to same code.
- Deriving a new calculated value (e.g. age = sys_date - d_o_b)
- Summarizing multiple rows of data.

The loading phase loads the transformed data into the data warehouse so that it can be used for various analytical purposes. Various reporting and analytical tools can be applied to data warehouse. Once data is loaded into data warehouse it cannot be updated. Loading is time consuming process so it is being done very few times.

A good ETL tool should be able to communicate with many different relational databases and read the various file formats used throughout an organization. ETL tools have started to migrate into Enterprise Application Integration, or even Enterprise Service Bus, systems that now cover much more than just the extraction, transformation and loading of data. Many ETL vendors

now have data profiling, data quality and metadata capabilities. ETL Data flow diagram is shown in figure18.

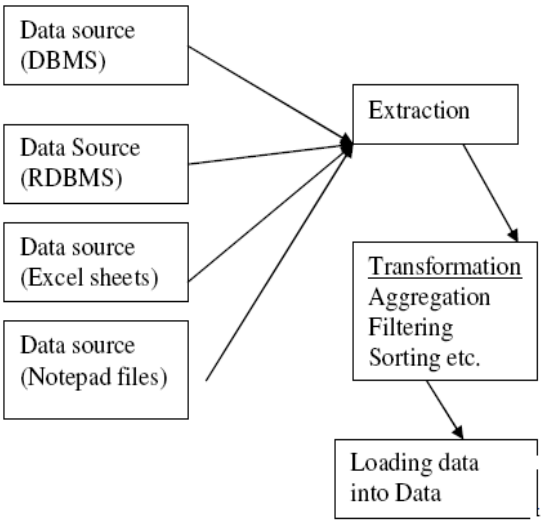

Figure 18. ETL Data flow diagram

3) Human resource management: TM techniques are also used to manage human resources strategically, mainly with applications aiming at analyzing staff's opinions, monitoring the level of employee satisfaction, as well as reading and storing CVs for the selection of new personnel. In the context of human resources management, the TM techniques are often utilized to monitor the state of health of a company by means of the systematic analysis of informal documents.

B. Text Mining Applications in Customer Relationship Management and Market analysis

Text Mining is widely used in field of Customer relationship Management and Market Analysis. Following are its few applications in these areas.

1) Customer Relationship Management (CRM): In CRM [4] domain the most widespread applications are related to the management of the contents of clients' messages. This kind of analysis often aims at automatically rerouting specific requests to the appropriate service or at supplying immediate answers to the most frequently asked questions. Services research has emerged as a green field area for application of advances in computer science and IT.

CRM practices, particularly contact centers (call centers) in our context, have emerged as hotbeds for application of innovations in the areas of knowledge management, analytics, and data mining. Unstructured text documents produced from a variety of sources in today contact centers have exploded in terms of the sheer volume generated. Companies are increasingly looking to understand and analyze this content to derive operational and business insights. The customer, the end consumer of products and services, is receiving increased attention.

Analytics and business intelligence (BI) applications revolving around the customer has led to emergence of areas like customer experience management, customer relationship management, and customer service quality. These are becoming critical to competitive growth, and sometimes even, survival. Applications with such customer focus are most evident in services companies, especially CRM practices and contact centers.

We focus on the particular application of C-Sat [12] (Customer Satisfaction) analysis in contact centers depicted in Figure 19.

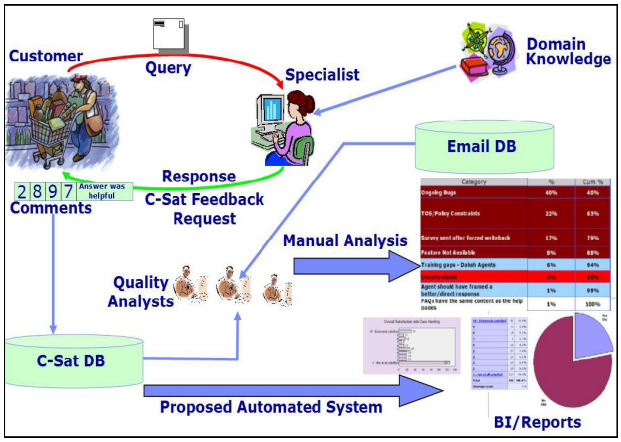

Figure 19. C-Sat analysis setting in contact centers and industrial CRM practices

Customers interacting with contact centers, typically fill out feedback surveys with free form text comments which are directly indicative of their (dis)satisfaction with products, services, or interactions. Currently, only manual analysis is done, if at all, on a small sample of available data. Various problem buckets identified by the analysis can directly lead to actions like improving a particular agent's accent, or imparting product training for other agents. Text classification can help automate this making it consistent and exhaustive.

2) Market Analysis (MA): Market Analysis, instead, uses TM mainly to analyze competitors and/or monitor customers' opinions to identify new potential customers, as well as to determine the companies' image through the analysis of press reviews and other relevant sources. For many companies tele-marketing and e-mail activity represents one of the main sources for acquiring new customers. The TM instrument makes it possible to present also more complex market scenarios.

Traditional marketing had a positive impact due to technology over the past few decades. Database technologies transformed storing information such as customers, partners, demographics, and preferences for making marketing decisions. In the 90s, the whole world saw economy boom due to improvements and innovation in various IT-related fields. The amount of web pages ameliorated during dot-com era. Search engines were found to crawl web pages to throw out useful information from the heaps. Marketing professionals used search engines, and databases as a part of competitive analyses.

Data mining technology helped extract useful information and find nuggets from various databases. Data warehouses turned out to be successful for numerical information, but failed when it came to textual information. The 21st century has taken us beyond the limited amount of information on the web. This is good in one way that more information would provide greater awareness, and better knowledge. In reality, it turns out to be not that good because too much of information leads to redundancy. The knowledge of marketing information is available on the web by means of industry white papers, academic publications relating to markets, trade journals, market news articles, reviews, and even public opinions when it comes down to customer requirements.

Text mining technology could help marketing professionals use this information for finding nuggets. Market Analysis includes following things:

- Where are the data sources for analysis?

Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies .

- Target marketing

Find clusters of "model" customers who share the same characteristics: interest, income level, spending habits, etc.

- Determine customer purchasing patterns over time

Conversion of single to a joint bank account: marriage, etc.

- Cross-market analysis

1. Associations/co-relations between product sales
2. Prediction based on the association information

- Finance planning and asset evaluation

1. Cash flow analysis and prediction
2. Contingent claim analysis to evaluate assets
3. Cross-sectional and time series analysis (financial ratio, trend analysis, etc.)

- Resource planning:

1. Summarize and compare the resources and spending competition:
2. Monitor competitors and market directions
3. Group customers into classes and a class-based pricing procedure
4. Set pricing strategy in a highly competitive market

It is instructive to divvy up the text-mining market [30] by the type of customer. Applying a bit of Bayesian reasoning, early buyers (as text mining is a relatively new technology) will prove a good indication of market directions. Nonetheless, text mining is testing new directions as the technology has made possible several new applications. The text-mining market is relatively new and not yet rigidly defined. Growth potential is huge given the ever-increasing volumes of textual information being produced and consumed by industry and government. While the market is being targeted by established software vendors, which are extending existing product lines and existing functional and industry-specific offerings, several pure-play vendors have done quite well and entry barriers for innovative start-ups are still not high.

C. Text Mining Applications in Technology watch

The technological monitoring [4], which analyses the characteristics of existing technologies, as well as identifying emerging technologies, is characterized by two elements: the capacity to identify in a non-ordinary way what already exists and that is consolidated and the capacity to identify what is already available, identifying through its potentiality, application fields and relationships with the existing technology.

Powerful Text Mining techniques now exist to identify the relevant Science & Technology [29] literatures, and extract the required information from these literatures efficiently. These techniques have been developed, to:

1) Substantially enhance the retrieval of useful information from global Science & Technology databases ;
2) Identify the technology infrastructure (authors, journals, organizations) of a technical domain;
3) Identify experts for innovation-enhancing technical workshops and review panels;
4) Develop site visitation strategies for assessment of prolific organizations globally;
5) Generate technical taxonomies (classification schemes) with human-based and computer-based clustering methods;
6) Estimate global levels of emphasis in targeted technical areas ;
7) Provide roadmaps for tracking myriad research impacts across time and applications areas.

Text mining has also been used or proposed for discovery and innovation from disjoint and disparate literatures. This application has the potential to serve as a cornerstone for credible technology forecasting, and help predict the technology directions of global military and commercial adversaries.

D. Text Mining Applications in Natural Language Processing and Multilingual Aspects

Text Mining is widely used in field of Natural Language Processing and Multilingual Aspects. Following are its few applications in these areas:

1) Questioning in Natural Language: The most important case of application of the linguistic competences developed in the TM context is the construction of websites that support systems of questioning in natural language. The need to make sites cater as much as possible for the needs of customers who are not necessarily expert in computers or web search is common also to those companies that have an important part of their business on the web.

The system architecture [31] of an automatic question answering system is shown in figure 20. In the user interface, users can question using natural language and then process automatic word segmentation. For the automatic question answering system, the user's question is for a specific course. The proposed question answering system is based on a specific course. Therefore, it is easy to extract keywords from the results of automatic word segmentation. The ontology-based knowledge base defines the concepts and the relationship between the concepts in the field of curriculum. The concepts have

been clearly defined and given the computer understandable semantics. With this domain ontology, the system can expand the keywords, increasing the search area for the problem, improve the system's recall rate. Then, the system uses the expanded key words to query in the FAQ base and return the handled answers to users.

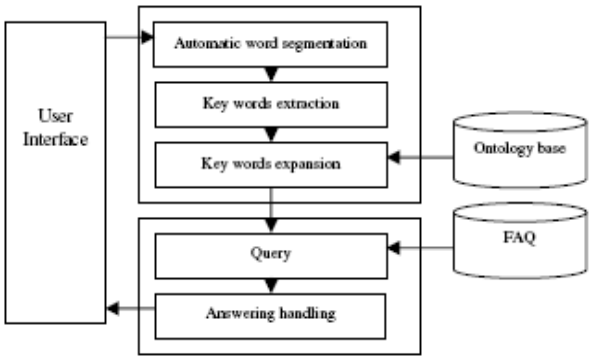

Figure 20. question answering system architecture

2) Multilingual Applications of Natural Language Processing: In NLP, Text Mining applications are also quite frequent and they are characterized by multi-linguism [4]. Use of Text Mining techniques to identify and analyze web pages published in different languages, is one of its example.

When working on a multilingual speech recognition system [32], a good deal of attention must be paid to the languages to be recognized by the system. In this application, a recognition system for Italian and German is built, thus the properties of both languages are of importance. From the acoustic point of view, the Italian language presents a significantly smaller number of phones with respect to German - e.g. 5 vowels in Italian versus 25 in German. Moreover, recognition experiments on Italian large vocabulary dictation conducted at IRST showed that only minor improvements are achieved with context-dependent (CD) phone models with respect to context-independent (CI) ones.

German is characterized by a higher variety of flexions and cases, a large use of compounding words, and the usage of capital and small letters to specify role of words. All these features heavily reflect on the vocabulary size and on out-of-vocabulary rate, that are in general higher for German. For the German language, it can be said, that pronunciation and lexicon strongly depend on the region. South Tyrolean German uses different words and pronunciation rules than standard German. Moreover, the land register experts have either Italian or German mother language and may thus have an accent whenever they enter data in a non-native language. Therefore, the recognition system must not only cope with dialectal variations, but also with a certain amount of accent by the speaker. The architecture for SpeeData data entry module is shown in figure21. This comprises four modules, namely the central manager (CM), the user interface (UI), the data base interface (DBI) and the speech recognizer *(SR).*

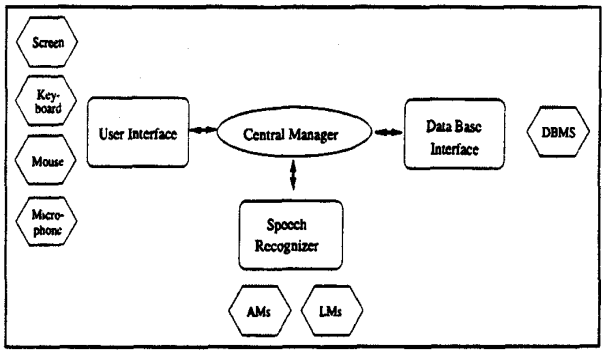

Figure 21. System architecture for the SpeeData data entry module

IV. DIFFERENCE BETWEEN TEXT MINING AND DATA MINING

The difference between regular data mining [2] and text mining is that in text mining the patterns are extracted from natural language text rather than from structured databases of facts. One application of text mining is in, bioinformatics where details of experimental results can be automatically extracted from a large corpus of text and then processed computationally. Text-mining techniques have been used in information retrieval systems as a tool to help users narrow their queries and to help them explore other contextually related subjects.

Text Mining seems to be an extension of the better known Data Mining. Data Mining is a technique that analyses billions of numbers to extract the statistics and trends emerging from a company's data. This kind of analysis has been successfully applied in business situations as well as for military, social, government needs. But, only about 20% of the data on intranets and on the World Wide Web are numbers - the rest is text. The information contained in the text (about 80% of the data) is invisible to the data mining programs that analyze the information flow in corporations.

Text mining tries to apply these same techniques of Data mining to unstructured text databases. To do so, it relies heavily on technology from the sciences of Natural Language Processing (NLP), and Machine Learning to automatically collect statistics and infer structure and meaning in otherwise unstructured text. The usual approach involves identifying and extracting key features from the text that can be used as the data and dimensions for analysis. This process is called feature extraction, is a crucial step in text mining.

Text mining is a comprehensive technique. It relates to data mining, computer language, information searching, natural language comprehension, and knowledge management. Text mining uses data mining techniques in text sets to find out connotative knowledge. Its object type is not only structural data but also semi-structural data or non-structural data. The mining results are not only general situation of one text document but also classification and clustering of text sets.

V. WEB MINING: KNOWLEDGE DISCOVERY FROM WEB

Buried in the enormous, heterogeneous and distributed information on the web is knowledge with great potential value. With the rapid development of the web, it is urgent and important to provide users with tools for efficient and effective resource discovery and knowledge discovery on the web. Although the web search engine assists in resource discovery, it is far from satisfying for its poor precision. Moreover, the target of the web search engine is only to discover resource on the web. As far as knowledge discovery is concerned, it is not equal to at all even with high precision. Therefore, the research and development of new technology further than resource discovery is needed.

Data mining is used to identify valid, novel, potentially useful and ultimately understandable pattern from data collection in database community. However, there is little work that deals with unstructured and heterogeneous information on the Web. Web mining [5],[8],[9] is the activity of identifying patterns implied in large document collection. Web mining is an integrated technology in which several research fields are involved, such as data mining, computational linguistics, statistics, informatics and so on. Different researchers from different communities disagree with each other on what web mining is exactly. Since web mining derives from data mining. Its definition is similar to the well-known definition of data mining.

Nevertheless, Web mining has many unique characteristics compared with data mining.
Firstly, the source of Web mining are web documents[10]. Web mining is to explore interesting information and potential patterns from the contents of web page, the information of accessing the web page linkages and resources of e-commerce by using techniques of data mining, which can help people extract knowledge, improve web sites design, and develop e-commerce better.

Using Knowledge Discovery in Database (KDD) [6], where the fundamental step is Data Mining, knowledge workers can obtain important strategic information for their business. KDD has deeply transformed the methods to interrogate traditional databases, where data are in structured form, by automatically finding new and unknown patterns in huge quantity of data. However, structured data represent only a little part of the overall organization knowledge. Knowledge is incorporated in textual documents. The amount of unstructured information in this form, accessible through the web, the intranets, the news groups etc is enormously increased in last years. In this scenario the development of techniques and instruments of Knowledge Extraction, that are able to manage the knowledge contained in electronic textual documents, is a necessary task. This is possible through a KDD process based on Text Mining.

A particular Text Mining approach is based on clustering techniques used to group documents according to their content. Knowledge discovery from texts (KDTs) involves discovering interesting unseen patterns in text databases, establishing which kind of knowledge should be acquired and how its novelty and interestingness should be evaluated is still a matter of research.

The metrics commonly used come from *data mining* (DM) or *knowledge discovery in databases* (KDDs) techniques and were developed for structured databases. However, they cannot be immediately applied to text data. Despite the large amount of research over the last few years, only few research efforts worldwide have realized the need for high-level representations (i.e., not just keywords), for taking advantage of linguistic knowledge and for the specific purpose of producing and assessing the unseen knowledge. The rest of the effort has concentrated on doing text mining from an information retrieval (IR) perspective and so both representation (keyword based) and data analysis are restricted. The most sophisticated approaches to text mining or KDT are characterized by an intensive use of external electronic resources, which highly restrict the application of the unseen patterns to be discovered

VI. CONCLUSIONS

At last we conclude that, Text mining is also known as Text Data Mining or Knowledge-Discovery in Text (KDT) [6], refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value. Knowledge may be discovered from many sources of information, yet, unstructured texts remain the largest readily available source of knowledge.

The problem of Knowledge Discovery from Text (KDT) is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process. KDT plays an increasingly significant role in emerging applications, such as Text Understanding. Most of Text Mining operations are as follows:

- Feature Extraction.
- Text-base navigation.
- Search and Retrieval
- Categorization (Supervised Classification)
- Clustering (Unsupervised Classification)
- Summarization

Primary objective of the Feature Extraction operation is to identify facts and relations in text. Most of the times this includes distinguishing which noun phrase is a person, place, organization or other distinct object. Feature Extraction algorithms may use dictionaries to identify some terms and linguistic patterns to detect others. The Text-base navigation enables users to move about in a document collection by relating topics and

significant terms. It helps to identify key concepts and additionally presents some of the relationships between key concepts.

Search and Retrieval is used for searching internal documents collections. Its main characteristic is the various text search options. After indexing which is the first step, a wide range of text search options may be utilized. These include basic search options such as Boolean (and/or/not), wildcard, segment, numeric range, etc. as well as some advance search capabilities, such as relevancy-ranked natural language searching, fuzzy search, concept search etc.

Categorization is the operation that we use, when we want to classify documents into predefined categories. Due to this, we are able to identify the main topics of a document collection. The categories are either pre-configured (by the programmer) or left for the user to specify.

A Cluster is a group of related documents, and Clustering [7]is the operation of grouping documents on the basis of some similarity measure, automatically without having to pre-specify categories. The most common Clustering algorithms that are used are hierarchical, binary relational, and fuzzy. Hierarchical clustering creates a tree with all documents in the root node and a single document in each leaf node. The intervening nodes have several documents and become more and more specialized as they get closer to the leaf nodes. It is very useful when we are exploring a new document collection and want to get an overview of the collection. The most important factor in a Clustering algorithm is the similarity measure. All Clustering algorithms are based on similarity measures.

Summarization is the operation that reduces the amount of text in a document while still keeping its key meaning. With this operation the user usually is allowed to define a number of parameters, including the number of sentences to extract or a percentage of the total text to extract.

Trend Analysis is used for identifying trends in documents collected over a period of time. Trends can be used, for example to discover that, a company is shifting interests from one domain to another.

In Attribute Analysis, given a set of documents, identify relationships between attributes (features that have been extracted from the documents) such as the presence of one pattern implies the presence of another pattern.

Visualization utilizes feature extraction and key term indexing in order to build a graphical representation of the document collection. This approach supports the user in identifying quickly the main topics or concepts by their importance on the representation. Additionally, it is easy to discover the location of specific documents in a graphical document representation. The ability to visualize large text data sets lets users quickly explore the semantic relationships that exist in a large collection of documents. In-formation visualization for text mining typically involves producing a 2D or 3D representation that exposes selected types of semantic patterns in a document collection.

To visualize text documents, we first need to convey to the analyst the underlying relation-ships between documents in a geometrically intuitive manner—we must preserve certain Visualization characteristics for the rendering to be meaningful. For example, documents that are close to each other in content should also be geometrically close to each other. Additionally, the analyst should be able to provide his or her own insight in determining what it means for documents to have close content or similar meaning.

For users to provide their own insight, they have to access the meaning of the Visualization. That is, users must be able to interpret the rendered data so that the topic-document relations are clearly defined.

REFERENCES

[1] [Berry Michael W., (2004), "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC, 24-43.

[2] Navathe, Shamkant B., and Elmasri Ramez, (2000), "*Data Warehousing And Data Mining*", in "*Fundamentals of Database Systems*", Pearson Education pvt Inc, Singapore, 841-872.

[3] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, (2005), "*Tapping into the Power of Text Mining*", Journal of ACM, Blacksburg.

[4] Sergio Bolasco , Alessio Canzonetti , Francesca Della Ratta-Rinald and Bhupesh K. Singh, (2002), "Understanding Text Mining:a Pragmatic Approach", Roam, Italy.

[5] Liu Lizhen, and Chen Junjie, China (2002), " Research of Web Mining", Proceedings of the 4th World Congress on Intelligent Control and Automation, IEEE, 2333-2337.

[6] Haralampos Karanikas and Babis Theodoulidis Manchester, (2001), "Knowledge Discovery in Text and Text Mining Software", Centre for Research in Information Management, UK

[7] Liritano S. and Ruffolo M., (2001), " *Managing the Knowledge Contained in Electronic Documents: a Clustering Method for Text Mining*", IEEE, 454-458 , Italy.

[8] Brin S., and Page L.(1998), " *The anatomy of a large-scale hyper textual Web search engine*", Computer Networks and ISDN Systems, 30(1-7): 107-117.

[9] Kleinberg J.M., (1999), "Authoritative sources in hyperlinked environment", Journal of ACM, Vol.46, No.5, 604-632.

[10] Dean J. and Henzinger M.R. (1999), "Finding related pages in the world wide web", Computer Networks, 31(11-16):1467-1479.

[11] N. Kanya and S. Geetha† (2007), "Information Extraction: A Text Mining Approach", IET-UK International Conference on Information and Communication Technology in Electrical Sciences, IEEE, Dr. M.G.R. University, Chennai, Tamil Nadu, India,1111-1118.

[12] Shantanu Godbole, and Shourya Roy, India (2008), "Text to Intelligence: Building and Deploying a Text Mining Solution in the Services Industry for Customer Satisfaction Analysis", IEEE, 441-448.

[13] Sungjick Lee and Han-joon Kim (2008), "News Keyword Extraction for Topic Tracking", Fourth International

Conference on Networked Computing and Advanced Information Management, IEEE, Koria,554-559.
[14] Joe Carthy and Michael Sherwood-Smith (2002), "Lexical chanins for topic tracking", International Conference, IEEE SMC WP1M1, Ireland.
[15] Wang Xiaowei, JiangLongbin, MaJialin and Jiangyan (2008), "Use of NER Information for Improved Topic Tracking", Eighth International Conference on Intelligent Systems Design and Applications, IEEE computer society, Shenyang, 165-170.
[16] Farshad Kyoomarsi ,Hamid Khosravi ,Esfandiar Eslami ,Pooya Khosravyan Dehkordy and Asghar Tajoddin (2008), "Optimizing Text Summarization Based on Fuzzy Logic", Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE computer society, 347-352.
[17] Fang Chen, Kesong Han and Guilin Chen (2008), "An approach to sentence selection based text summarization", Proceedings of IEEE TENCON02, 489- 493.
[18] Setu Madhavi Namburu, Haiying Tu, Jianhui Luo and Krishna R. Pattipati (2005), "Experiments on Supervised Learning Algorithms for Text Categorization", International Conference , IEEE computer society, 1-8.
[19] Guihua Wen, Gan Chen, and Lijun Jiang (2006), "Performing Text Categorization on Manifold", 2006 IEEE International Conference on Systems, Man, and Cybernetics, Taipei, Taiwan , IEEE , 3872-3877.
[20] JIAN-SUO XU (2007), "TCBPLK: A new method of text categorization ", Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong,, IEEE , 3889-3892.
[21] Ming Zhao, Jianli Wang and Guanjun Fan (2008), "Research on Application of Improved Text Cluster Algorithm in intelligent QA system ", Proceedings of the Second International Conference on Genetic and Evolutionary Computing, China,, IEEE Computer Society , 463-466.
[22] XiQuan Yang, DiNa Guo, XueYa Cao and JianYuan Zhou (2008), "Research on Ontology-based Text Clustering", Third International Workshop on Semantic Media Adaptation and Personalization, China,, IEEE Computer Society , 141-146.
[23] Zhou Ning, Wu Jiaxin, Wang Bing and Zhang Shaolong (2008), "A Visualization Model for Information Resources Management", 12th International Conference Information Visualisation, China**,** IEEE, 57- 62.
[24] Jignashu Parikh and M. Narasimha Murty (2002),"Adapting Question Answering Techniques to the Web", Proceedings of the Language Engineering Conference, India**,** IEEE computer society.
[25] Emilio Sanchis, Davide Buscaldi, Sergio Grau, Lluis Hurtado and David Griol (2006)," SPOKEN QA BASED ON A PASSAGE RETRIEVAL ENGINE", Proceedings of IEEE international confernce, Spain, 62-65.
[26] Hwee-Leng Ong, Ah-Hwee Tan, Jamie Ng, Hong Pan and Qiu-Xiang Li.(2001),"FOCI : Flexible Organizer for Competitive Intelligence", In Proceedings, Tenth International Conference on Information and Knowledge Management (CIKM'01), pp. 523-525, Atlanta, USA, 5-10.
[27] Li Gao, Elizabeth Chang, and Song Han (2005)," Powerful Tool to Expand Business Intelligence: Text Mining", PROCEEDINGS OF WORLD ACADEMY OF SCIENCE, ENGINEERING AND TECHNOLOGY VOLUME 8, 110-115.
[28] Rajender Singh Chhillar (2008)," Extraction Transformation Loading –A Road to Data warehouse", 2nd National Conference Mathematical Techniques: Emerging Paradigms for Electronics and IT Industries,India, 384-388.
[29] RONALD NEIL KOSTOFF (2003),"TEXT MINING FOR GLOBAL TECHNOLOGY WATCH", article, OFFICE OF NAVAL RESEARCH, Quincy St. Arlington, 1-27.
[30] Seth Grimes (2005),"The developing text mining market", white paper, Text Mining Summit05 Alta Plana Corporation**,**,Boston, 1-12.
[31] Wang Bo and Li Yunqing (2008)," Research on the Design of the Ontology-based Automatic Question Answering System", International Conference on Computer Science and Software Engineering, IEEE, Nanchang, China, 871-874.
[32] U. Ackermann, B. Angelini, F. Brugnara, M. Federico ,D. Giuliani, R. Gretter, G. Lazzari and H. Niemann ,"SpeeData: Multilingual Spoken Data Entry", International Conference, IEEE, Trento, Italy., 2211 - 2214.
[33] Dion H. Goh and Rebecca P. Ang (2007)," An introduction to association rule mining: An application in counseling and help seeking behavior of adolescents", Journal of Behavior Research Methods39 (2), *Singapore,* 259-266
[34] Pak Chung Wong, Paul Whitney and Jim Thomas," Visualizing Association Rules for Text Mining", ", International Conference, Pacific Northwest National Laboratory, *USA,* 1-5

Vishal Gupta is Lecturer in Computer Science & Engineering Department at University Institute of Engineering & Technology, Panjab university Chandigarh. He has done MTech. in computer science & engineering from Punjabi University Patiala in 2005. He was among university toppers. He secured 82% Marks in MTech. Vishal did his BTech. in CSE from Govt. Engineering College Ferozepur in 2003. He is also pursuing his PhD in Computer Sc & Engg. Vishal is devoting his research work in field of Natural Language processing. He has developed a number of research projects in field of NLP including synonyms detection, automatic question answering and text summarization etc. One of his research paper on Punjabi language text processing was awarded as best research paper by Dr. V. Raja Raman at an International Conference at Panipat. He is also a merit holder in 10^{th} and 12^{th} classes of Punjab School education board. in professional societies. The photograph is placed at the top left of the biography. Personal hobbies will be deleted from the biography.

Gurpreet Singh Lehal, professor, received undergraduate degree in Mathematics in 1988 from Panjab University, Chandigarh, India, and Post Graduate degree in Computer Science in 1995 from Thapar Institute of Engineering & Technology, Patiala, India and Ph. D. degree in Computer Science from Punjabi University, Patiala, in 2002. He joined Thapar Corporate R&D Centre, Patiala, India, in 1988 and later in 1995 he joined Department of Computer Science at Punjabi University, Patiala. He is actively involved both in teaching and

research. His current areas of research are- Natural Language Processing and Optical Character recognition. He has published more than 25 research papers in various international and national journals and refereed conferences. He has been actively involved in technical development of Punjabi and has to his credit the first Gurmukhi OCR, Punjabi word processor with spell checker and various transliteration software. He was the chief coordinator of the project "Resource Centre for Indian Language Technology Solutions- Punjabi", funded by the Ministry of Information Technology as well as the coordinator of the Special Assistance Programme (SAP-DRS) of the University Grants Commission (UGC), India. He was also awarded a research project by the International Development Research Centre (IDRC) Canada for Shahmukhi to Gurmukhi Transliteration Solution for Networking.