Volume 4, Issue 11, November 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

The Acumen and Acuity of Data Mining and Text Mining

Anjali Jivani, Jait Purohit, Kaushal Patel
Comp Science & Engineering MSU
Baroda, India

Abstract— In this paper we have focused a variety of techniques in kaleidoscopic way, approaches and different areas of the research which are helpful and marked as the important field of Data Mining. Large and renowned organizations may generate large volumes of data. Corporate decision makers require access from all such sources and take strategic decision. The data warehouse is used in the significant business value by improving the effectiveness of important decision-making. To analyze, manage and make a decision of such type of huge amount of data we need techniques called the data mining which will transforming in many fields. This paper includes more number of applications of the data mining and also focuses scope of the data mining which will be helpful in the further research.

Apart from this we have focused on Text Mining, which is an important step of Knowledge Discovery process. It is used to extract hidden information from not-structured or semi-structured data. In this paper, our basic focus is to study the concept of Text Mining and various techniques. Here, we are able to determine how to mine the Plain as well as Structured Text.

We have also covered the concept of Natural Language Processing (NLP). NLP is an engineering discipline which uses computer to do useful things using human language. A glimpse of Computational Linguistics (CL) is also described. The difference between NLP and CL is juxtaposed below.

Keywords— Data Mining, Text Mining, Structured Text Mining, Unstructured Text Mining, Computational Linguistics, Natural Language Processing.

I. INTRODUCTION

Data mining, popularly known as Knowledge Discovery in Databases (KDD), it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. It is actually the process of finding the hidden information from the repositories.

Text mining is also known as text data mining, which refers to the process of extracting high-quality information from text. Text mining includes the process of structuring the input text like parsing and other successive insertion into a database. Text Mining deals with the structured data, evaluates them and finally produces the result. Text mining includes text categorization, text clustering, sentiment analysis, document summarization, and entity relation modelling. Text mining is a process that has a set of algorithms for converting unstructured text into structured data objects and the quantitative methods used to analyse these data objects.

Computational linguistics (CL) is a discipline between linguistics and computer science which deals with the computational aspects of the human language faculty. It belongs to the cognitive sciences and also deals with the field of artificial intelligence (AI), a branch of computer science aiming at computational models of human cognition.

Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation.

II. THE DATA MINING TASK

The data mining tasks are of different types depending on the use of data mining result the data mining tasks are classified as:

A. Exploratory Data Analysis

In the repositories vast amount of information is available. This data mining task will serve the two purposes (i) Without the knowledge for what the customer is searching, then (ii) It analyses the data.

These techniques are interactive and visual to the customer.

B. Descriptive Modelling

It describe all the data, it includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables.

C. Predictive Modelling

This model permits the value of one variable to be predicted from the known values of other variables.

D. Discovering Patterns and Rules

This task is primarily used to find the hidden pattern as well as to discover the pattern in the cluster. In a cluster a number of patterns of different size and clusters are available .The aim of this task is "how best we will detect the patterns" .This can be accomplished by using rule induction and many more techniques in the data mining algorithm like(K-Means/K-Medoids). These are called the clustering algorithm.

E. Retrieval by Content

The primary objective of this task is to find the data sets frequently used for audio/video along-with images. It is finding pattern similar to the pattern of interest in the data set.

III. TYPES OF DATA MINING SYSTEM

Data mining systems can be categorized according to various criteria the classification is as follows:

A. Classification of data mining systems according to the type of data source mined

In an organization a huge amount of data is available where we need to classify these data but these are available most of times in a similar fashion. We need to classify these data according to its type (maybe audio/video, text format etc).

B. Classification of data mining systems according to the data model

There are so many number of data mining models (Relational data model, Object Model, Object Oriented data Model, Hierarchical data Model/W data model) are available and each and every model we are using the different data. According to these data model the data mining system classify the data in the model.

C. Classification of data mining systems according to the kind of knowledge discovered

This classification based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

D. Classification of data mining systems according to mining techniques used

This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc. The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options, and offer different degrees of user interaction.

IV. DATA MINING LIFE CYCLE

The life cycle of a data mining project consists of six phases. The sequence of the phases is not rigid. Moving back and forth between different phases is always allowed and required. It depends on the outcome of each phase. The main phases are shown in Fig. 1:

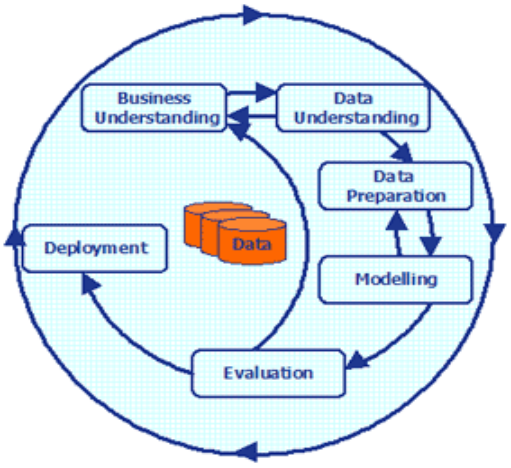

Fig. 1 Data Mining Life Cycle

A. Business Understanding

This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Data Understanding

It deals with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

B. Data Preparation

In this stage, it collects all the different data sets and construct the varieties of the activities basing on the initial raw data.

C. Modelling

In this phase, various modelling techniques are selected and applied and their parameters are calibrated to optimal values.

D. Evaluation

In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.

E. Deployment

The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

V. DATA MINING APPLICATIONS

In this section, we have focused some of the applications of data mining and its techniques are analysed respectively.

A. Data Mining Applications in Healthcare

Data mining applications in health can have tremendous potential and usefulness. However, the success of healthcare data mining hinges on the availability of clean healthcare data. In this respect, it is crucial that the healthcare industry look into how data can be better captured, stored, prepared and mined. Possible directions contain the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications. As healthcare data are not limited to just quantitative data (e.g., doctor's notes or clinical records), it is necessary to also explore the use of text mining to expand the scope and nature of what healthcare data mining can currently do. This is specially used to mixed all the data and then mining the text. It is also useful to look into how images (e.g., MRI scans) can be brought into healthcare data mining applications. It is noted that progress has been made in these areas.

B. Data mining is used for market basket analysis

Data mining technique is used in MBA(Market Basket Analysis).When the customer want to buy some products then this technique helps us finding the associations between different items that the customer put in their shopping buckets. Here the discovery of such associations that promotes the business technique. In this way the retailers uses the data mining technique so that they can identify that which customers intension (buying the different pattern).In this way this technique is used for profits of the business and also helps to purchase the related items.

C. Data mining is used an emerging trends in the education system in the whole world

In Indian culture most of the parents are uneducated .The main aim of in Indian government is the quality education not for quantity. But the day by day the education systems are changed and in the 21st century a huge number of universalities are established by the order of UGC. As the numbers of universities are established side by side, each and every day a millennium of students are enrols across the country. With huge number of higher education aspirants, we believe that data mining technology can help bridging knowledge gap in higher educational systems. The hidden patterns, associations, and anomalies that are discovered by data mining techniques from educational data can improve decision making processes in higher educational systems. This improvement can bring advantages such as maximizing educational system efficiency, decreasing student's drop-out rate, and increasing student's promotion rate, increasing student's retention rate in, increasing student's transition rate, increasing educational improvement ratio, increasing Student's success, increasing student's learning outcome, and reducing the cost of system processes. In this current era we are using the KDD and the data mining tools for extracting the knowledge this knowledge can be used for improving the quality of education .The decisions tree classification is used in this type of applications.

D. Data mining is now used in many different areas in manufacturing engineering

When we retrieve the data from manufacturing system then the customer is to use these data for different purposes like to find the errors in the data ,to enhance the design methodology ,to make the good quality of the data ,how best the data can be supported for making the decision . But most of time the data can be first analysed then after find the hidden patterns which will be control the manufacturing process which will further enhance the quality of the products .Since the

importance of data mining in manufacturing has clearly increased over the last 20 years, it is now appropriate to critically review its history and Application.

E. In Medical Science

In medical science there is large scope for application of data mining. Diagnosis of diesis, health care, patient profiling and history generation etc. are the few examples. Mammography is the method used in breast cancer detection. Radiologists face lot of difficulties in detection of tumours that's why CAM (Computer Aided Methods) could helps to the medical staff. So that they can produce the good quality of the result detection. The neural networks with back-propagation and association rule mining used for tumour classification in mammograms. The data mining effectively used in the diagnosis of lung abnormality that may be cancerous or benign. The data mining algorithms significantly reduce patient's risks and diagnosis costs. Using the prediction algorithms the observed prediction accuracy was 100% for 91.3% cases. The use of data mining in health care is the widely used application of data mining. The medical data is complex and difficult to analyse. A REMINDS (Reliable Extraction and Meaningful Inference from Non-structured Data) system integrates the structured and unstructured clinical data in patient records to automatically create high quality structured clinical data. To adopt the high quality technique, we can mine the existing patient records to support guidelines and give necessary help to improve patient care.

F. The Intrusion Detection in the Network

The intrusion detection in the Network is very difficult and needs a very close watch on the data traffic. The intrusion detection plays an essential role in computer security. The classification method of data mining is used to classify the network traffic normal traffic or abnormal traffic. If any TCP header does not belong to any of the existing TCP header clusters, then it can be considered as anomaly.

G. Sports data Mining

The data mining and its technique is used for an application of Sports centre. Data mining is not only use in the business purposes but also it used in the sports .In the world, a huge number of games are available where each and every day the national and international games are to be scheduled, where a huge number of data's are to be maintained .The data mining tools are applied to give the information as and when we required. The open source data mining tools like WEKA and RAPID MINER frequently used for sport. This means that users can run their data through one of the built-in algorithms, see what results come out, and then run it through a different algorithm to see if anything different stands out. As these programs' are available in the form of open source in nature, that's why the users are frequently to modify the source code, so that other can get the updated information. In the sports world the vast amounts of statistics are collected for each player, team, game, and season. In the game sports the data's are available in the form of statistical form where data mining can be used and discover the patterns, these patterns are often used to predict the future forecast. Data mining can be used for scouting, prediction of performance, selection of players, coaching and training and for the strategy planning. The data mining techniques are used to determine the best or the most optimal squad to represent a team in a team, sport in a season, tour or game.

H. The Intelligence Agencies

The Intelligence Agencies collect and analyse information to investigate terrorist activities. One challenge to law enforcement and intelligent agencies is the difficulty of analysing large volume of data involve in criminal and terrorist activities. Nowadays the intelligence agency are using the sophisticated data mining algorithms which makes it easy, to handle the very large data bases databases for organizations. The different data mining techniques are used in crime data mining. Though the organization's have used large data bases but data mining helps us to generate the different types of information in the organization like personal details of the persons along with, vehicle details .In data mining the Clustering techniques is used (Association rule mining) for the different objects(like persons, organizations, vehicles etc.) in crime records. Not only data mining detects but also analyses the crime data. The classification technique is also used to detect email spamming and also find person who has given the mail. String comparator is used to detect deceptive information in criminal record.

I. E-commerce

E-commerce is also the most prospective domain for data mining. It is ideal because many of the ingredients required for successful data mining are easily available: data records are plentiful, electronic collection provides reliable data, insight can easily be turned into action, and return on investment can be measured. The integration of e-commerce and data mining significantly improve the results and guide the users in generating knowledge and making correct business decisions. This integration effectively solves several major problems associated with horizontal data mining tools including the enormous effort required in pre-processing of the data before it can be used for mining, and making the results of mining actionable.

VI. THE SCOPE OF DATA MINING

Data mining derives its name from the similarities by searching for valuable business information in a large database for example, finding linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore. Both processes require either shifting through an immense amount of material, or intelligently searching it to find exactly

where the value resides. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms by default, and identifying segments of a population likely to respond similarly to given events. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

1) *Artificial neural networks:* Non-linear predictive models that learn through training and resemble biological neural networks in structure.
2) *Decision trees:* Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
3) *Genetic algorithms:* Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
4) *Nearest neighbour method:* A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbour technique.
5) *Rule induction:* The extraction of useful if-then rules from data based on statistical significance. Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms.

VII. TEXT MINING

A. Introduction to text mining

The Text mining processes unstructured information, extracts meaningful numeric indices from the text, and makes the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. Information can be extracted from the summarized words of the documents, so the words can be analysed and also the similarities between words and documents can be determined or how they are related to other variables in the data-mining project. Basically, text mining converts text into numbers which can then be included in other analyses such as predictive data mining projects, clustering etc. Text mining is also known as text data mining, which refers the process of deriving high-quality information from text. High-quality information is derived through the statistical pattern learning. Text mining includes the process of structuring the input text like parsing and other successive insertion into a database. TM derives patterns within the structured data, evaluates them and finally produces the output. Text mining takes account of text categorization, text clustering, sentiment analysis, document summarization, and entity relation modelling. Text mining is a process that employs a set of algorithms for converting unstructured text into structured data objects and the quantitative methods used to analyse these data objects.

B. Applications of text mining

Here are various applications of Text mining like automatic processing of messages and emails. For example, it is possible to "filter" out automatically "junk email" based on certain terms; such messages can automatically be discarded. Such automatic systems for classifying electronic messages can also be useful in applications where messages need to be routed automatically to the most appropriate department. Another application is analysing warranty or insurance claims, diagnostic interviews. In some business domains, the majority of information is collected in textual form. For example, warranty claims or initial medical (patient) interviews can be summarized in brief narratives, or when you take your automobile to a service station for repairs, typically, the attendant will write some notes about the problems that you report and what you believe needs to be fixed. Increasingly, those notes are collected electronically, so those types of narratives are readily available for input into text mining algorithms. Analysing open-ended survey responses. Survey questionnaires typically shoot two broad types of questions: open-ended and closed-ended. Closed-ended questions give a discrete set of responses from which to choose. Such types of responses are easily quantified and analysed while open-ended questions allow the respondent to answer a question in his own words. Such types of unstructured responses often provide richer and more valued information than closed-ended questions and are an important source of insight since they can generate information that was not anticipated. Despite their added value, researchers often prefer to avoid including open-ended questions in their surveys because of the tedious task of reading and coding responses, a time-consuming and expensive task especially when one has more than a few hundred written responses.

C. Differences between various terminologies of text mining

1) *Text Mining vs. Data Mining:* In Text Mining, patterns are extracted from natural language text but in Data Mining patterns are extracted from databases.
2) *Text Mining vs. Web Mining:* In Text Mining, the input is free unstructured text, but in Web Mining web sources are structured.

VIII. METHODS OF MINING TEXT

A. Mining Plain Text

This section describes the major ways in which text is mined when the input is plain natural language, rather than partially-structured Web documents. We begin with problems that involve extracting information for human consumption. Here are the various techniques which mine the plain text like text summarization, document retrieval, Information retrieval, Assessing document similarity and Text categorization.

B. Text summarization

A text summarizer produces a compressed representation of its input, which specifies human consumption. It also contains individual documents or groups of documents. Text compression is a related area but the output of text summarization is specific to be human-readable. The output of text compression algorithms is definitely not human-readable and it is also not actionable, it only supports decompression, that is, automatic reconstruction of the original text. Summarization differs from many other forms of text mining in that there are people, namely professional abstractors, who are skilled in the art of producing summaries and carry out the task as part of their professional life.

C. Document Retrieval

Document retrieval is the task of identifying and returning the most relevant documents. Traditional libraries provide catalogues that allow users to identify documents based on resources which consist of metadata. Metadata is a highly structured document for summary, and successful methodologies have been developed for manually extracting metadata and for identifying relevant documents based on it, methodologies that are widely taught in library school. Automatic extraction of metadata (e.g. subjects, language, author, key-phrases) is a prime application of text mining techniques. The idea is to index every individual word in the document collection. It specifies many effective and popular document retrieval techniques.

D. Information retrieval

Information retrieval is considered as an extension to document retrieval where the documents that are returned are processed to condense or extract the particular information sought by the user. Thus document retrieval is followed by a text summarization stage that focuses on the query posed by the user, or an information extraction stage. The modularity of documents may be adjusted so that each individual subsection or paragraph comprises a unit in its own right, in an attempt to focus results on individual nuggets of information rather than lengthy documents.

E. Assessing document similarity

Many text mining problems involve assessing the similarity between different documents; for example, assigning documents to pre-defined categories and grouping documents into natural clusters. These are the basic problems in data mining too, and have been a focus for research in text mining, perhaps because the success of different techniques can be evaluated and compared using standard, objective, measures of success.

F. Text categorization

Text categorization is the assignment of natural language documents to predefined categories according to their content. The set of categories is often called a "controlled vocabulary." Document categorization is a long-standing traditional technique for information retrieval in libraries, where subjects rival authors as the predominant gateway to library contents—although they are far harder to assign objectively than authorship. Automatic text categorization has many practical applications, including indexing for document retrieval, automatically extracting metadata, word sense disambiguation by detecting the topics a document covers, and organizing and maintaining large catalogues of Web resources. As in other areas of text mining, until the 1990s text categorization was dominated by ad hoc techniques of "knowledge engineering" that sought to elicit categorization rules from human experts and code them into a system that could apply them automatically to new documents. Since then—and particularly in the research community—the dominant approach has been to use techniques of machine learning to infer categories automatically from a training set of pre-classified documents. Indeed, text categorization is a hot topic in machine learning today. The pre-defined categories are symbolic labels with no additional semantics. When classifying a document, no information is used except for the document's content itself. Some tasks constrain documents to a single category, whereas in others each document may have many categories. Sometimes category labelling is probabilistic rather than deterministic, or the objective is to rank the categories by their estimated relevance to a particular document. Sometimes documents are processed one by one, with a given set of classes; alternatively there may be a single class—perhaps a new one that has been added to the set—and the task is to determine which documents it contains. Many machine learning techniques have been used for text categorization.

G. Document clusterisation

Imaging a database of customer records, where each record represents a customer's attributes. These can include identifiers such as name and address, demographic information such as gender and age, and financial attributes such as income and revenue spent. Clustering is an automated process to group related records together. Related records are grouped together on the basis of having similar values for attributes. This approach of segmenting the database via clustering analysis is often used as an exploratory technique because it is not necessary for the end-user/analyst to specify

ahead of time how records should be related together. In fact, the objective of the analysis is often to discover segments or clusters, and then examine the attributes and values that define the clusters or segments. As such, interesting and surprising ways of grouping customers together can become apparent, and this in turn can be used to drive marketing and promotion strategies to target specific types of customers. There are a variety of algorithms used for clustering, but they all share the property of iteratively assigning records to a cluster, calculating a measure (usually similarity, and/or distinctiveness), and re-assigning records to clusters until the calculated measures don't change much indicating that the process has converged to stable segments. Records within a cluster are more similar to each other, and more different from records that are in other clusters. Depending on the particular implementation, there are a variety of measures of similarity that are used (e.g. based on spatial distance, based on statistical variability, or even adaptations of Condorcet values used in voting schemes), but the overall goal is for the approach to converge to groups of related records.

IX. MINING STRUCTURED TEXT

Much of the text that we have on the Internet contains explicit structural mark-up and differs from traditional plain text. Some mark-up is internal and indicates document structure or format; some is external and gives explicit hypertext links between documents. These information sources give additional benefits for mining Web documents. Both sources of information are extremely noisy: they involve arbitrary and unpredictable choices by individual page designers. However, these disadvantages are offset by the total amount of data that is available, which is relatively unbiased because it is aggregated over many different information providers. Thus "Web mining" is emerging as a new subfield, similar to text mining but taking advantage of the extra information available in Web documents, particularly hyperlinks—and even capitalizing on the existence of topic directories in the Web itself to improve results. We briefly review three techniques for mining structured text. The first, wrapper induction, uses internal mark-up information to increase the effectiveness of text mining in marked-up documents. The remaining two, document clustering and determining the "authority" of Web documents, capitalize on the external mark-up information that is present in hypertext in the form of explicit links to other documents.

A. Wrapper Induction

Internet resources that contain relational data—telephone directories, product catalogs, etc.—use Formatting mark-up to clearly present the information they contain to users. However, with standard HTML, it is quite difficult to extract data from such resources in an automatic way. The XML mark-up language is designed to overcome these problems by encouraging page authors to mark their content in a way that reflects document structure at a detailed level; but it is not clear to what extent users will be prepared to share the structure of their documents fully in XML, and even if they do, huge numbers of legacy pages abound. Many software systems use external online resources by hand-coding simple parsing modules, commonly called "wrappers," to analyse the page structure and extract the requisite information. This is a kind of text mining, but one that depends on the input having a fixed, predetermined structure from which information can be extracted algorithmically. Given that this assumption is satisfied, the information extraction problem is relatively trivial. But this is rarely the case. Page structures vary; errors that are insignificant to human readers throw automatic extraction procedures off completely; Web sites evolve. There is a strong case for automatic induction of wrappers to reduce these problems when small changes occur, and to make it easier to produce new sets of extraction rules when structures change completely.

B. Document clustering with links

Document clustering techniques are based on the documents' textual similarity. However, the hyperlink structure of Web documents, encapsulated in the "link graph" in which nodes are Web pages and links are hyperlinks between them, can be used as a different basis for clustering. Many standard graph clustering and partitioning techniques are applicable. Link-based clustering schemes typically use factors such as: The number of hyperlinks that must be followed to travel in the Web from one document to the other; the number of common ancestors of the two documents, weighted by their ancestry distance and the number of common descendents of the documents, similarly weighted. These can be combined into an overall similarity measure between documents. In practice, a textual similarity measure is usually incorporated as well, to yield a hybrid clustering scheme that takes account of both the documents' content and their linkage structure. The overall similarity may then be determined as the weighted sum of four factors. Such a measure will be sensitive to the characteristics of the documents and their linkage structure, and given the number of parameters involved there is considerable scope for tuning to maximize performance on particular data sets.

C. Determining "authority" of Web documents

The Web's linkage structure is a valuable source of information that reflects the popularity, sometimes interpreted as "importance," "authority" or "status," of Web pages. For each page, a numeric rank is computed. The basic premise is that highly-ranked pages are ones that are cited, or pointed to, by many other pages. Consideration is also given to (a) the rank of the citing page, to reflect the fact that a citation by a highly-ranked page is a better indication of quality than one from a lesser page, and (b) the number of out-links from the citing page, to prevent a highly ranked page from artificially magnifying its influence simply by containing a large number of pointers. This leads to a simple algebraic equation to determine the rank of each member of a set of hyperlinked pages. Complications arise from the fact that some links are "broken" in that they lead to nonexistent pages, and from the fact that the Web is not fully connected; these are easily overcome. Such techniques are widely used by search engines (e.g. Google) to determine how to sort the hits associated

with any given query. They provide a social measure of status that relates to standard techniques developed by social scientists for measuring and analysing social networks.

X. NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is the computerized approach to analysing text that is based on both a set of theories and a set of technologies. And, being a very active area of research and development, there is not a single agreed-upon definition that would satisfy everyone, but there are some aspects, which would be part of any knowledgeable person's definition.

Natural Language Processing is a theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

A. Goal

The goal of NLP as stated above is "to accomplish human-like language processing". The choice of the word 'processing' is very deliberate, and should not be replaced with 'understanding'. For although the field of NLP was originally referred to as Natural Language Understanding (NLU) in the early days of AI, it is well agreed today that while the goal of NLP is true NLU, that goal has not yet been accomplished. A full NLU System would be able to:

1) Paraphrase an input text
2) Translate the text into another language
3) Answer questions about the contents of the text
4) Draw inferences from the text

While NLP has made serious inroads into accomplishing goals 1 to 3, the fact that NLP systems cannot, of themselves, draw inferences from text, NLU still remains the goal of NLP. There are more practical goals for NLP, many related to the particular application for which it is being utilized. For example, an NLP-based IR system has the goal of providing more precise, complete information in response to a user's real information need. The goal of the NLP system here is to represent the true meaning and intent of the user's query, which can be expressed as naturally in everyday language as if they were speaking to a reference librarian. Also, the contents of the documents that are being searched will be represented at all their levels of meaning so that a true match between need and response can be found, no matter how either are expressed in their surface form.

B. Divisions

While the entire field is referred to as Natural Language Processing, there are in fact two distinct focuses – language processing and language generation. The first of these refers to the analysis of language for the purpose of producing a meaningful representation, while the latter refers to the production of language from a representation. The task of Natural Language Processing is equivalent to the role of reader/listener, while the task of Natural Language Generation is that of the writer/speaker. While much of the theory and technology are shared by these two divisions, Natural Language Generation also requires a planning capability. That is, the generation system requires a plan or model of the goal of the interaction in order to decide what the system should generate at each point in an interaction. We will focus on the task of natural language analysis, as this is most relevant to Library and Information Science. Another distinction is traditionally made between language understanding and speech understanding. Speech understanding starts with, and speech generation ends with, oral language and therefore rely on the additional fields of acoustics and phonology. Speech understanding focuses on how the 'sounds' of language as picked up by the system in the form of acoustical waves are transcribed into recognizable morphemes and words. Once in this form, the same levels of processing which are utilized on written text are utilized. All of these levels, including the phonology level, will be covered in Section 2; however,
The emphasis throughout will be on language in the written form.

XI. A GLIMPSE INTO COMPUTATIONAL LINGUISTICS

"Human knowledge is expressed in language. So computational linguistics is very important." –Mark Steedman, ACL Presidential Address (2007).

Computational linguistics is the scientific and engineering discipline concerned with understanding written and spoken language from a computational perspective, and building artefacts that usefully process and produce language, either in bulk or in a dialogue setting. To the extent that language is a mirror of mind, a computational understanding of language also provides insight into thinking and intelligence. And since language is our most natural and most versatile means of communication, linguistically competent computers would greatly facilitate our interaction with machines and software of all sorts, and put at our fingertips, in ways that truly meet our needs, the vast textual and other resources of the internet.

XII. DIFFERENCE BETWEEN COMPUTATIONALLINGUISTICS AND NATURAL LANGUAGE PROCESSING

Table I difference between CL and NLP

Computational Linguistics	**Natural Language Processing**

• Computational linguistics is a scientific discipline that studies linguistic processes from a computational perspective. 1) Language comprehension (computational psycho-linguistics) 2) Language production 3) Language acquisition •Computational Linguistics seeks to study language using computers and corpora. •The field of Computational Linguistics (CL) aims to model aspects of the human language faculty using formal computational models (of both symbolic and statistical varieties), with the aim of understanding the nature of language as a phenomenon.	• Natural language processing is an engineering discipline that uses computers to do useful things with language. 1) Information retrieval 2) Topic detection and document clustering 3) Document summarisation 4) Sentiment analysis 5) Machine translation 6) Speech recognition •NLP seeks to do useful things using human language. •The field of NLP is applied in a huge variety of language technologies, NLP researchers focus on cross-cutting techniques.

XIII. CONCLUSION

To conclude our study paper in a nut-shell, we would like to say that data mining is an important field and deals with mining of large data from huge data warehouses and other data repositories. Data mining is gradually increasing in each and every field and its applications are also varied and kaleidoscopic. Text mining is used to extract hidden information from unstructured data given in various formats. Natural Language Processing helps to probe various data and text using human language whereas Computational Linguistics studies language using computers and logistics. This field is a sheer ocean of research for those inclined towards it..

REFERENCES

[1] S.R.Pande, S.S Sambare, V. M Thakre, "Data clustering using data mining techniques", International journal of advance research in computer and communication engineering Vol1. 1, Issue 8, October-2012.
[2] Thomas Miller, "Data and Text Mining", Pearson Education, 2008.
[3] Steven Bird, Ewan Klein & Edward Loper, "Natural Language Processing with Python", with O'Reilly.
[4] Ritu Arora, "Text Mining: Classification and Clustering" University of Alabama, Birmingham.
[5] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques" Elsevier (2001).
[6] Daniel T. Larose, "Data Mining Methods and Models" Wiley-Interscience.
[7] David Hand, Heikki Mannila, Padhraic Symth, "Principles of Data Mining", PHI.
[8] Ronen Feldman, James Sanger, "The Text Mining Handbook", Cambridge University Press, 2006.
[9] Manu Kochady "Text Mining Application Programming", Thomson India Edition, 2006.
[10] Hearst, M. Untangling Text Data Mining. In the proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999.

