ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

AN ANALYSIS ON TEXT MINING -TEXT RETRIEVAL AND TEXT EXTRACTION

Umajancy.S[1], Dr. Antony Selvadoss Thanamani[2]

Research Scholar, Dr. Mahalingam Centre for Research and Development, NGM College, Pollachi, India[1]

Associate Professor and Head, Dr. Mahalingam Centre for Research and Development, NGM College, Pollachi, India [2]

Abstract: Text Mining is the analysis of data contained in natural language text. Text Mining works by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a data base and analyzed with traditional data mining techniques. Data stored in text database is mostly semi structured i.e., it is neither completely unstructured nor completely structured. Information retrieval techniques such as text indexing have been developed to handle the unstructured documents. The related task of Information Extraction (IE) is about locating specific items in natural language documents. This article analyses the various techniques related to text retrieval and text extraction.

Keywords: Text Mining, Information retrieval, Information Extraction, Natural Language Processing**.**

I. INTRODUCTION

Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. Text databases are rapidly growing due to the increasing amount of information available in electronic form, such as electronic publications, various kinds of electronic documents, e-mail, and the World Wide Web. Nowadays most of the information in government, industry, business, and other institutions are stored electronically, in the form of text databases.

Data stored in most text databases are semi structured data in that they are neither completely unstructured nor completely structured. For example, a document may contain a few structured fields, such as title, authors, publication date, and category, and so on, but also contain some largely unstructured text components, such as abstract and contents. There has been a great deal of Studies on the modelling and implementation of semi structured data in recent database research. Moreover, information retrieval techniques, such as text indexing methods, have been developed to handle unstructured documents. Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data. Typically, only a small fraction of the many available Documents will be relevant to a given individual user. Without knowing what could be in the documents, it is difficult to formulate effective queries for analysing and extracting useful information from the data. Users need tools to compare different documents, rank the importance and relevance of the documents, or find patterns and trends across multiple documents. Thus, text mining has become an increasingly popular and essential theme in data mining.

II. TEXT MINING

Text mining or knowledge discovery from text (KDT) — for the first time mentioned in Feldman et al. [6] — deals with the machine supported analysis of text. It uses techniques from information retrieval, Information extraction, as well as natural language processing (NLP) and connects them with the algorithms and methods of KDD, data mining, machine learning and statistics. Thus, one selects a similar procedure as with the KDD process, whereby not data in general, but text documents are in focus of the analysis. From this, new questions for the used data mining methods arise. One problem is that now has to deal with problems of — from the data modelling perspective— unstructured data sets.

Information Extraction

The first approach assumes that text mining essentially corresponds to information extraction

Text Data Mining

Text mining can be also defined — similar to data mining — as the application of algorithms and methods from the field's machine learning and statistics to texts with the goal of finding useful patterns. For this purpose it is necessary to pre-process the texts accordingly. Many authors use information extraction methods, natural language processing or some simple pre-processing steps in order to extract data from texts. To the extracted data then data mining algorithms can be applied [4, 7].

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

III. TECHNIQUES OF TEXT MINING

A. Information Retrieval

Information retrieval is a field that has been developing in parallel with database systems for many years. Unlike the field of database systems, which has focused on query and transaction processing of structured data, information retrieval is concerned with the organization and retrieval of information from a large number of text-based documents. Since information retrieval and database systems each handle different kinds of data, some database system problems are usually not present in information retrieval systems, such as concurrency control, recovery, transaction management, and update. Also, some common information retrieval problems are usually not encountered in traditional database systems, such as unstructured documents, approximate search based on keywords, and the notion of relevance. Due to the abundance of text information, information retrieval has found many applications. There exist many information retrieval systems, such as on-line library catalog systems, on-line document management systems, and the more recently developed Web search engines. A typical information retrieval problem is to locate relevant documents in a document collection based on a user's query, which is often some keywords describing an information need, although it could also be an example relevant document. In such a search problem, a user takes the initiative to "pull" the relevant information out from the collection; this is most appropriate when a user has some ad hoc information need, such as finding information to buy a used car. When a user has a long-term information need , a retrieval system may also take the initiative to "push" any newly arrived information item to a user if the item is judged as being relevant to the user's information need. Such an information access process is called information filtering, and the corresponding systems are often called filtering systems or recommender systems. From a technical viewpoint, however, search and filtering share many common techniques.

i. Measures for Text Retrieval

The set of documents relevant to a query be denoted as *{*Relevant*}*, and the set of documents retrieved be denoted as {Retrieved}. The set of documents that are both relevant and retrieved is denoted as {Relevant} ∩ {Retrieved}, as shown in the Venn diagram of Figure 1. There are two basic measures for assessing the quality of text retrieval.

Precision: This is the percentage of retrieved documents that are in fact relevant to the query. It is formally defined as

$$\text{Precision} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|}$$

Recall: This is the percentage of documents that are relevant to the query and were, in fact, retrieved. It is formally defined as

$$\text{Recall} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Relevant}\}|}$$

An information retrieval system often needs to trade off recall for precision or vice versa. One commonly used trade-off is the F-score, which is defined as the harmonic mean of recall and precision

$$\text{F-score} = \frac{\text{Recall} \times \text{Precision}}{(\text{Recall}+\text{Precision})/2}$$

The harmonic mean discourages a system that sacrifices one measure for another too drastically.

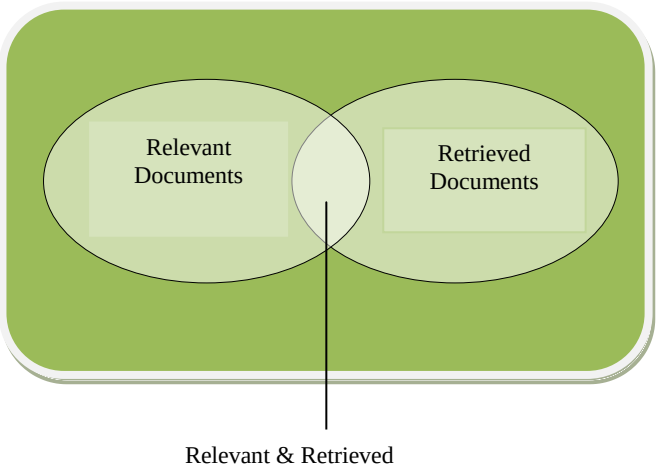

Fig 1. Relationship between the set of Relevant Documents and the set of Retrieved Documents.

Precision, recall, and F-score are the basic measures of a retrieved set of documents. These three measures are not directly useful for comparing two ranked lists of documents because they are not sensitive to the internal ranking of the documents in a retrieved set. In order to measure the quality of a ranked list of documents, it is common to compute an average of precisions at all the ranks where a new relevant document is returned. It is also common to plot a graph of precisions at many different levels of recall; a higher curve represents a better-quality information retrieval system [3].

Recall measures the quantity of relevant results returned by a search, meanwhile precision is the measure of the quality of the results returned. Recall is the ratio of relevant results returned divided by all relevant results. Precision is the number of relevant results returned divided by the total number of results returned.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

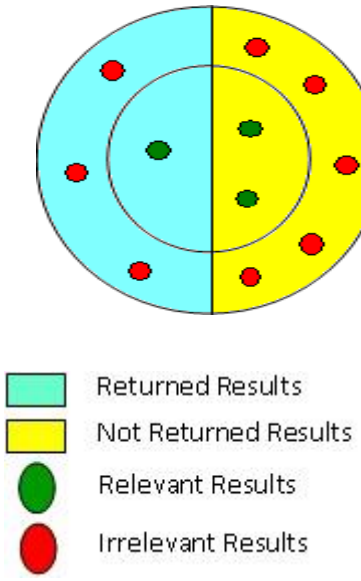

Fig 2. Represents a low-precision, low-recall search as described in the text

The figure above represents a low-precision, low-recall search. In the diagram the red and green dots represent the total population of potential search results for a given search. Red dots represent irrelevant results, and green dots represent relevant results. Relevancy is indicated by the proximity of search results to the center of the inner circle. Of all possible results shown, those that were actually returned by the search are shown on a light-blue background. In the example only one relevant result of three possible relevant results was returned, so the recall is a very low ratio of 1/3 or 33%. The precision for the example is a very low 1/4 or 25%, since only one of the four results returned was relevant [8].

B. Text Indexing

When dealing with a small number of documents, it is possible for the full-text-search engine to directly scan the contents of the documents with each query, a strategy called "serial scanning." This is what some rudimentary tools, such as Grep, do when searching.

However, when the number of documents to search is potentially large, or the quantity of search queries to perform is substantial, the problem of full-text search is often divided into two tasks: indexing and searching. The indexing stage will scan the text of all the documents and build a list of search terms (often called an index, but more correctly named a concordance). In the search stage, when performing a specific query, only the index is referenced, rather than the text of the original documents.

The indexer will make an entry in the index for each term or word found in a document, and possibly note its relative position within the document. Usually the indexer will ignore stop words (such as "the" and "and") that are both common and insufficiently meaningful to be useful in searching. Some indexers also employ language-specific stemming on the words being indexed. For example, the words "drives", "drove", and "driven" will be recorded in the index under the single concept word "drive." [2].

C. Information Extraction (IE)

IE involves directly with text mining process by extracting useful information from the texts. IE deals with the extraction of specified entities, events and relationships from unrestricted text sources. IE can be described as the creation of a structured representation of selected information drawn from texts. In IE natural language texts are mapped to be predefine, structured representation, or templates, which, when it is filled, represent an extract of key information from the original text [10], [11].

The goal is to find specific data or information in natural language texts. Therefore the IE task is defined by its input and its extraction target. The input can be unstructured documents like free texts that are written in natural language or the semi-structured documents that are pervasive on the Web, such as tables or itemized and enumerated lists. Using IE approach, events, facts and entities are extracted and stored into a structured database.

Then data mining techniques can be applied to the data for discovering new knowledge. Unlike information retrieval (IR), which concerns how to identify relevant documents from a document collection, IE produces structured data ready for post-processing, which is crucial to many text mining applications.

Figure 3 illustrates how IE can play a part in a knowledge mining process. Furthermore, IE allows for mining the actual information present within the text, rather than the limited set of tags associated to the documents. The work of [11], [4], have presented how information extraction is used for text mining.
According to [12] and [13] typical IE are developed using the following three steps:-

- Text pre-processing; whose level ranges from text segmentation into sentences and sentences into tokens, and from tokens into full syntactic analysis;
- Rule selection; the extraction rules are associated with triggers (e.g. keywords), the text is scanned to identify the triggering items and the corresponding rules are selected;

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

- Rule application, which checks the conditions of the selected rules and fill in the form according to the conclusions of the matching rules.

Furthermore [14] and [15] emphasized that information extraction is based on understanding of the structure and meaning of the natural language in which documents are written, and the goal of information extraction is to accumulate semantic information from text.

Technically, extracting information from texts requires two pieces of knowledge: lexical knowledge and linguistic grammars. Using the knowledge we are able to describe the syntax and semantic of the text [16]. A common approach to information extraction is to use patterns which match against text and identify items of interest. Patterns are applied to texts which have undergone various levels of linguistic analysis, such as phrase chunking [17] and full syntactic parsing [18]. The approaches may use different definition of what constitutes a valid pattern. For example, [19] use subject-verb-object tuples derived from a dependency parse, followed by [20] uses patterns which match certain grammatical categories, mainly nouns and verbs, in phrase chunked text.

Reference [21] reported in identifying the parts of a person name through analysis of name structure. For example, the name Doctor Paul R. Smith is composed of a person title, a first name, a middle name, and a surname. It is presented as a pre-processing step for entity recognition and for the resolution of co references to help determine, for instance, that John F. Kennedy and President Kennedy is the same person, while John F. Kennedy and Caroline Kennedy are two distinct persons.

Research work in [22] applied IE for detecting events in text. Event detection consists of detecting temporal entities in conjunction with other entities. For example, conferences are usually made up of four parts: one conference name, one location, and two dates (e.g., name: "AAAI," location: "Boston," start date: "July 16th 2006," end date: "July 20th 2006"). A person birth or death is a person name and date pair (e.g., name: "John Lennon," date: "December 8th, 1980"). Smith used event detection to draw maps where war locations and dates are identified [23].

D. Natural Language Processing (NLP)

NLP is a technology that concerns with natural language generation (NLG) and natural language understanding (NLU). NLG uses some level of underlying linguistic representation of text, to make sure that the generated text is grammatically correct and fluent. Most NLG systems include a syntactic realiser to ensure that grammatical rules

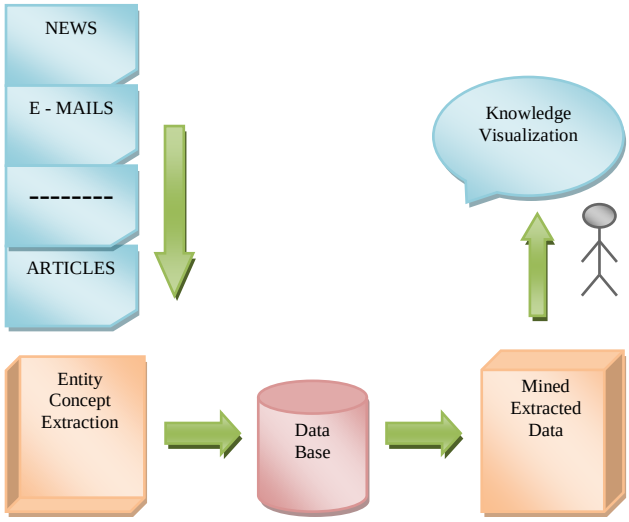

Fig 3 shows important entities are extracted and stored in a database. Data mining approach is used to mine the stored data. Hidden knowledge is then visualized.

such as subject-verb agreement are obeyed, and text planner to decide how to arrange sentences, paragraph, and other parts coherently. The most well known NLG application is machine translation system. The system analyses texts from a source language into grammatical or conceptual representations and then generates corresponding texts in the target language. NLU is a system that computes the meaning representation, essentially restricting the discussion to the domain of computational linguistic. NLU consists of at least of one the following components; tokenization, morphological or lexical analysis, syntactic analysis and semantic analysis. In tokenization, a sentence is segmented into a list of tokens. The token represents a word or a special symbol such an exclamation mark. Morphological or lexical analysis is a process where each word is tagged with its part of speech.

The complexity arises in this process when it is possible to tag a word with more than one part of speech. Syntactic analysis is a process of assigning a syntactic structure or a parse tree, to a given natural language sentence. It determines, for instance, how a sentence is broken down into phrases, how the phrases are broken down into sub-phrases, and all the way down to the actual structure of the words used [9].

Semantic analysis is a process of translating a syntactic structure of a sentence into a semantic representation that is precise and unambiguous representation of the meaning expressed by the sentence. A semantic representation allows a system to perform an appropriate task in its application domain. The semantic representation is in a formally

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

specified language. The language has expressions for real world objects, events, concepts, their properties and relationships, and so on. Semantic interpretation can be conducted in two steps: context independent interpretation and context interpretation. Context independent interpretation concerns what words mean and how these meanings combine in sentences to form sentence meanings. Context interpretation concerns how the context affects the interpretation of the sentence. The context of the sentence includes the situation in which the sentence is used, the immediately preceding sentences, and so on.

IV. CONCLUSION

With the dramatic increase in online information in recent years, text mining at the intersection of data mining, natural language processing, machine learning, and information retrieval, is starting to gain increasing interest. Most of knowledge hidden in electronic media of an organization is encapsulated in documents. Acquiring this knowledge implies effective querying of the documents as well as the combination of information pieces from different textual sources (e.g.: the World Wide Web). Integrating a domain knowledge base with a text mining engine would boost its efficiency, especially in the information retrieval and information extraction phases**.**

REFERENCES

[1] R. Sagayam, S. Srinivasan, S. Roshni, "A Survey Of Text Mining: Retrieval Extraction And Indexing Techniques", International Journal Of Computational Engineering Research, Volume: 2, Issue: 5, ISSN: 2250-3005, September 2012.

[2] Capabilities of Full Text Search System, http://en.wikipedia.org/wiki/Full_text_search#False-positive_problem.

[3] R. Baeza-yates and B.Ribeiro-neto, modern Information Retrieval addition – Wesley, Boston, 1999.

[4] U. Nahm and R. Mooney, "Text mining with information extraction", In Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, 2002.

[5] Namita Gupta, "Text Mining For Information Retrieval", May 2011.

[6] R. Feldman and I. Dagan. Kdt - knowledge discovery in texts. In Proc. of the First Int. Conf. on Knowledge Discovery (KDD), pages 112–117, 1995.

[7] R. Gaizauskas, "An information extraction perspective on text mining: Tasks, technologies and prototype applications", http://www.itri.bton.ac.uk/projects/euromap/TextMiningEvent/Rob_Gaizauskas.pdf, 2003.

[8] Coles, Michael (2008). Pro Full-Text Search in SQL Server 2008 (Version 1 Ed.). Apress Publishing Company. ISBN 1-4302-1594-1.

[9] S.Jusoh and H.M. Alfawareh, "Natural language interface for online sales," in Proceedings of the International Conference on Intelligent and Advanced System (ICIAS2007). Malaysia: IEEE, November 2007, pp. 224– 228.

[10] R. Rao, "From unstructured data to actionable intelligence," in Proceedings of the IEEE Computer Society, 2003.

[11] H. Karanikas, C. Tjortjis, and B. Theodoulidis, "An approach to text mining using information extraction," in Proceedings of Workshop of Knowledge Management: Theory and Applications in Principles of Data Mining and Knowledge Discovery 4th European Conference, 2000.

[12] R. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, and M. Tyson, "FASTUS: A cascaded finite-state transducer for extraction information from natural language text," in Finite States Devices for Natural Language Processing, E. Roche and Y. Schabes, Eds., 1997, pp. 383– 406.

[13] J. Cowie and Y. Wilks, "Information extraction", New York, 2000.

[14] N. Singh, "The use of syntactic structure in relationship extraction," Master's thesis, MIT, 2004.

[15] R. Hale, "Text mining: Gettingmore value from literature resources," Drug Discovery Today, vol. 10, no. 6, pp. 377– 379, 2005.

[16] C. Nedellec and A. Nazarenko, "Ontologies and information extraction: A necessary symbiosis," in Ontology Learning from Text: Methods, Evaluation and Applications, P. Buitelaar, P. Comiano, and B. Magnin, Eds. IOS Press Publication, 2005.

[17] S. Soderland, "Learning information extraction rules for semi-structured and free text," Machine Learning, vol. 34, pp. 233–272, 1999.

[18] R. Gaizauskas, T. Wakao, K. Humphreys, H. Cunningham, and Y. Wilks, "Description of the lasie system as used for muc-6," in Proceedings of the Sixth Message Understanding Conference (MUC-6), 1996, pp. 207– 220.

[19] R. Yangarber, R. Grishman, P. Tapanainen, and S. Huttunen, "Automatic acquisition of domain knowledge for information extraction," in Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000, 2000, pp. 940–946.

[20] R. Jones, R. Ghani, T. Mitchell, and E. Riloff, "Active learning for information extraction with multiple view feature sets," in Proceedings of the 20th International Conference on Machine Learning (ICML 2003), August 21- 24 2003.

[21] U. Charniak, "Unsupervised learning of name structure from co reference data," in Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics, 2001, pp. 48–54.

[22] D. Smith, "Detecting and browsing events in unstructured text," in Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, 2002.

[23] Shaidah Jusoh, Hejab M. Alfawareh, "Techniques, Applications and Challenging Issue in Text Mining", International Journal of Computer Science Issues, Volume: 9, Number: 2, Issue: 6, November 2012, ISSN: 1694-0814.