International Journal of Information Technology and Management

Vol. V, Issue No. I, August-2013, ISSN 2249-4510

AN ANALYSIS ON TEXT MINING AND TEXT RETRIEVAL TECHNIQUES

AN INTERNATIONALLY INDEXED PEER REVIEWED & REFEREED JOURNAL

www.ignited.in

International Journal of Information Technology and Management
Vol. V, Issue No. I, August-2013, ISSN 2249-4510

An Analysis on Text Mining and Text Retrieval Techniques

A. Sindhu[1]* Dr. C. A. Kanabar[2]

[1]Laxmi Institute of Commerce and Computer Application (BBA-BCA) Sarigam

[2]Assistant Professor (Computer Science) Saurastra University, Rajkot

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, Text Retrieval, Techniques, etc.

----------♦----------

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 have been a great deal of studies on the modeling 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 (Sagayam, 2012). 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 analyzing 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 (Michael, 2008).

REVIEW OF LITERATURE:

Text mining or knowledge discovery from text (KDT) — for the first time mentioned in (Jusoh and Alfawareh, 2007) 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.

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 (Singh, 2004). 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 (Hale, 2005). From a technical viewpoint, however, search and filtering share many common techniques. Below we briefly discuss the major techniques in information retrieval with a focus on search techniques.

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 |{Retrieved{|/|{Retrieved}∩{Relevant{|=Precision
Recall: This is the percentage of documents that are relevant to the query and were, in fact, retrieved. It is formally defined as

$$\text{F-score} = \text{recall} \times \text{precision} / (\text{recall} + \text{precision}) / 2$$

An information retrieval system often needs to trade off recall for precision or vice versa. One commonly used tradeoff is the F-score, which is defined as the harmonic mean of recall and precision F-score =recall × precision/ (recall+ precision)/2

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

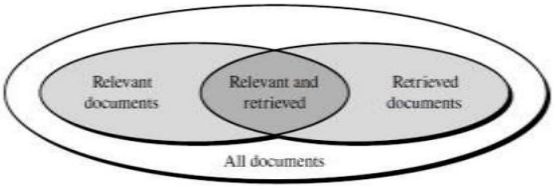

Figure 1. Relationship between the set of relevant documents and the set of retrieved documents

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 (Shaidah, Alfawareh (2012). For more details about these measures, readers may consult an information retrieval textbook, such as (Jusoh and Alfawareh, 2007).

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.

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 (Hearst, 2003). 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.

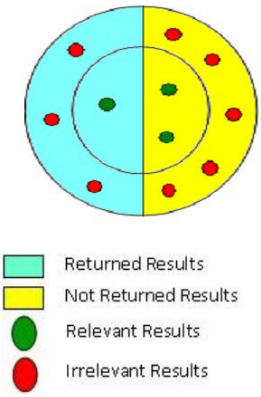

Fig 2. Represents a low-precision, low-recall search as described in the text

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.

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 (Sagayam, 2012). (Singh, 2004). have presented how information extraction is used for text mining. According to (Michael, 2008) and (Jusoh and Alfawareh, 2007). 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;
- Rule application, which checks the conditions of the selected rules and fill in the form according to the conclusions of the matching rules.

Furthermore (Singh, 2004) and (Hale, 2005). 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.

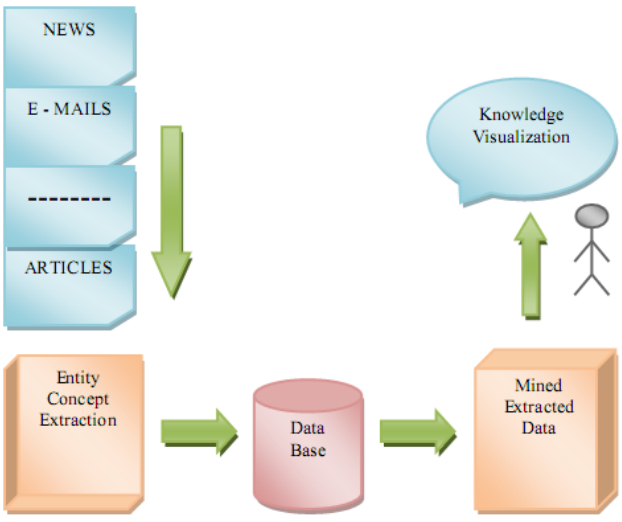

Fig 3 shows important entities are extracted and stored in a database. Data mining approach is used to mine the stored data.

Information Extraction: The general purpose of Knowledge Discovery is to "extract implicit, previously unknown, and potentially useful information from data". Information Extraction IE mainly deals with identifying words or feature terms from within a textual file. Feature terms can be defined as those which are directly related to the domain.

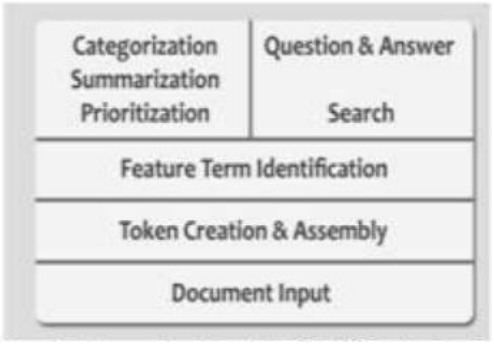

Figure 4. A layered model of the Text Mining Application These is the terms which can be recognized by the tool.

Stemming: Stemming refers to identifying the root of a certain word. There are basically two types of stemming techniques, one is inflectional and other is derivational. Derivational stemming can create a new word from an existing word, sometimes by simply changing grammatical category (for example, changing a noun to a verb). The type of stemming we were able to implement is called Inflectional Stemming. A commonly used algorithms is the 'Porter's Algorithm' for stemming. When the normalization is confined to regularizing grammatical variants such as singular/plural or past/present, it is referred to inflectional stemming. To minimize the effects of inflection and morphological variations of Words (stemming), our approach has pre-processed each word using a provided version of the Porter stemming algorithm with a few changes towards the end in which we have omitted some cases.

Domain dictionary: In order to develop tools of this sort, it is essential to provide them with a knowledge base. A collective set of the entire feature terms' is the Domain dictionary. The structure of the Domain dictionary which we implemented consisted of three levels in the hierarchy. Namely, Parent Category, Sub-category and word. Parent categories define the main category under which any sub-category or word falls. A parent category will be unique on its level in the hierarchy. Sub-categories will belong to a certain parent category and each subcategory will consist of all the words associated with it. As an example, consider the following Table 1 is an example that shows how we identify words which belong to the Parent Category 'Hardware' and Subcategory 'Input Devices'.

Table 1.Structure of the Domain Dictionary

Parent Category	Sub-Category	Words
Hardware	Data Storage	*Grabber*
	Input devices	*Light pen*
	Modems	*Joystick*
	Motherboards	*Contact image sensor*
	Networking	*Digital camera*

Exclusion List: A lot of words in a text file can be treated as unwanted noise. To eliminate these, we devised a separate file which includes all such words. These include words such as the, a, an, if, off, on etc.

CONCLUSION:

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). Discovering such hidden know ledge is an essential requirement for many corporations, due to its wide spectrum of applications. In this short survey, the notion of text mining has been introduced and several techniques available have been presented. Due to its novelty, there are many potential research areas in the field of Text Mining, which includes finding better intermediate forms for representing the outputs of information extraction, an XML document may be a good choice. Mining texts in different languages is a major problem, since text mining tools should be able to work with many languages and multilingual documents. Integrating a domain knowledge base with a text mining engine would boost its efficiency, especially in the information retrieval and information extraction phases.

REFERENCES:

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

M. A. Hearst (2003). What is text mining? http://www.sims.berkeley.edu/~hearst/text-mining.html.

N. Singh (2004). The use of syntactic structure in relationship extraction,‖ Master's thesis, MIT.

R. Hale (2005). Text mining: Gettingmore value from literature resources,‖ Drug Discovery Today, vol. 10, no. 6, pp. 377– 379.

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

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

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

Corresponding Author

A. Sindhu*

Laxmi Institute of Commerce and Computer Application (BBA-BCA) Sarigam

E-Mail – sindhubhilai@yahoo.com