A Survey Of Different Text Mining Techniques

Varsha C. Pande[1] and Dr. A.S. Khandelwal[2]

[1]*Department of Electronics & Comp. Sc, RTMNU, Nagpur, India*

[2]*Department of Computer Science, Hislop College, Nagpur, India*

ABSTRACT

Text mining is a technology that can work with unstructured or semi-structured data. It is a technology that can be used to find the meaningful information from natural language text using existing data in corporate databases by making unstructured text data available for analysis. There are many techniques for text mining. In this paper we describe the techniques, Information Extraction, Information retrieval, Query processing, Natural Language processing, Categorization, Clustering.

Keywords: Categorization, Clustering, Information Extraction, Information Retrieval, Natural Language Processing, Natural Language Text, Query Processing, Text mining.

Introduction

Text mining also known as text data mining (Sagayam, Srinivasan, Roshni, 2012) is a variation on a field called data mining, which is used to find interesting patterns from large databases. The data mining tools 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.

The phrase "text mining" (Witten) is basically used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract useful information. It discovers new pieces of knowledge from textual data. Basically it is used to combine countless pages of plain-language digitized text to find useful information that has been hiding in plain sight.

Most of the industries, government sectors, organizations and institutions data are stored in electronic form. These data are stored in text database format. Text database is semi structured format which contains many structured fields and few unstructured fields. For example students roll no, name, semester, class are the structured fields and Address, remarks are unstructured

fields in an institution. Text mining is essential for an organization because most of the information in the organizations is in text format.

Generally text mining involves following steps:

(1) Convert unstructured text inputs into structured database.

(2) Identify the patterns and trends from the structured data.

(3) Analyze and interpret the patterns and trends.

(4) Extracting the useful information from the text.

Text Mining Techniques

The main purpose of text mining techniques is to structure the text documents. The following are the important text mining techniques.

1. Information Extraction
2. Information retrieval
3. Natural Language processing
4. Query processing
5. Categorization
6. Clustering

Information Extraction (IE)

Information Extraction (Winter School) is a process of automatically extracting structured information from unstructured or semi-structured natural language text. Pattern matching is the final output of the extraction process. It is the type of database obtained by looking for predefined sequences in text. This involves defining the general form of the information that we are interested in as one or more templates, which are then used to guide the extraction process. IE systems rely heavily on the data generated by NLP systems.

The main objective of information extraction method is the extraction of specific information from text documents. These are stored in data base-like patterns and are then available for further use.

Functions performed by IE systems include:

Term analysis, which determines the terms appearing in a document. This is especially useful for documents that contain many complex multi-word terms, such as scientific research papers.

Named-entity recognition, which mainly specifies the names appearing in a document, such as names of people or organizations. Some systems are also able to recognize dates and expressions of time, quantities and associated units, percentages, and so on.

Fact extraction, which identifies and extract complex facts from documents. Such facts could be relationships between entities or events.

IE transforms a quantity of textual documents into a more structured database, the database assembled by an IE module then can be provided to the KDD module for promote mining of knowledge.

Information retrieval

Information Retrieval (IR) systems identify the documents in a collection which match a user's query. The most important application of information retrieval is search engine in World Wide Web, which identify those documents on the WWW that are important to a set of given words. The process of finding information according to the user's request is information retrieval. Normally, it refers to the automatic retrieval of documents. Information retrieval deals with crawling, indexing document and retrieving document. It is used to retrieve collection of significant pages from the set of pages in WWW. Database system contracts with query based structured data. Information Retrieval deals with query based on large amount of text documents. "Document" (Ramanathan, Meyyappan, 2013) is the standard term for an information holder. Information retrieval system used in online digital library, government online services and online document system and web search engines. For the large amount of text data the traditional information retrieval techniques are not enough. Text mining techniques such as text classification, text categorization, text summarization are powerful techniques to handle large amount of text data.

Since database and information retrieval systems each handle different types of data, some database system problems are generally not present in information retrieval systems, such as concurrency control, recovery, transaction management, and update. A document collection based on a user's query is typical information retrieval problem is to locate relevant documents, 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 adhoc 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.

Information retrieval is the finding of documents which contain answers to questions and not the finding of answers itself. In order to achieve this goal statistical measures and methods are used for the automatic processing of text data and comparison to the given question. Information retrieval in the broader sense deals with the entire range of information processing, from data retrieval to knowledge retrieval.

Natural Language processing

The general purpose of NLP is to achieve a better understanding of natural language by use of computers. The range of the assigned techniques reaches from the simple manipulation of strings to the automatic processing of natural language inquiries.

Natural language processing (NLP) is one of the oldest and most difficult problems in the field of artificial intelligence. It is the analysis of human language so that computers can understand natural languages as humans do. NLP (Natural Language Processing) technology is used to analyze text and cross reference data with large databases of background knowledge. The highly scalable text mining system is accessed by means of a web-based interface which allows simultaneous users to be located on different sites anywhere in the world.

The role of NLP (Ghosh, Roy, Bandyopadhyay, 2012) in text mining is to provide the systems in the information extraction phase with linguistic data that they need to perform their task. Often this is done by annotating documents with information such as sentence boundaries, part-of-speech tags and parsing results, which can then be read by the information extraction tools.

NLP (Ghosh, Roy, Bandyopadhyay, 2012) can perform some types of analysis with a high degree of success. Shallow parsers identify only the main grammatical elements in a sentence, such as noun phrases and verb phrases, whereas deep parsers generate a complete representation of the grammatical structure of a sentence. The role of NLP in text mining is to provide the systems in the information extraction phase (see below) with linguistic data that they need to perform their task. Often this is done by annotating documents with information like sentence boundaries, part-of-speech tags, parsing results, which can then be read by the information extraction tools.

Query processing

Once an inverted index is created for a document collection, a retrieval system can answer a keyword query quickly by looking up which documents contain the query keywords. Specifically, we will maintain a score accumulator for each document and update these accumulators as we go through each query term. For each query term, we will fetch all of the documents that match the term and increase their scores. When examples of relevant documents (Sagayam, Srinivasan, Roshni, 2012) are available, the system can learn from such examples to improve retrieval performance. This is called relevance feedback and has proven to be effective in improving retrieval performance. When we do not have such relevant examples, a system can assume the top few retrieved documents in some initial retrieval results to be relevant and extract more related keywords to expand a query. Such feedback is called pseudo-feedback or blind feedback and is essentially a process of mining useful keywords from the top retrieved documents. Pseudo-feedback also often leads to improved retrieval performance. One major limitation of many existing retrieval methods is that they are based on exact keyword matching. However, due to the complexity of natural languages, keyword based retrieval can encounter two major difficulties.

The first is the synonymy problem: two words with identical or similar meanings may have very different surface forms. For example, a user's query may use the word "automobile," but a relevant document may use "vehicle" instead of "automobile."

The second is the polysemy problem: the same keyword, such as mining, or Java, may mean different things in different contexts.

Clustering

The technique in which objects of logically similar properties are physically placed together in one class of objects and a single access to the disk makes the entire class available is clustering. This technique is used to group similar documents, but it differs from categorization 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.

There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset.

Clustering (Gupta, Lehal, 2009) methods can be divided into two categories based on the cluster structure which they produce

1. The non-hierarchical method
2. The hierarchical method

The non-hierarchical methods divide a dataset of *N* objects into *M* clusters, with or without overlap. These methods are divided into partitioning methods, in which the classes are mutually exclusive, and the less common clumping methods, in which overlap is allowed. Each object is a member of the cluster with which it is most similar; however the threshold of similarity has to be defined.

The hierarchical methods produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains. The hierarchical methods can be further divided into agglomerative or divisive methods.

In word relativity-based clustering (WRBC) method (Gupta, Lehal, 2009), text clustering process contains four main parts: text reprocessing, word relativity computation, word clustering and text classification. See in following Figure.

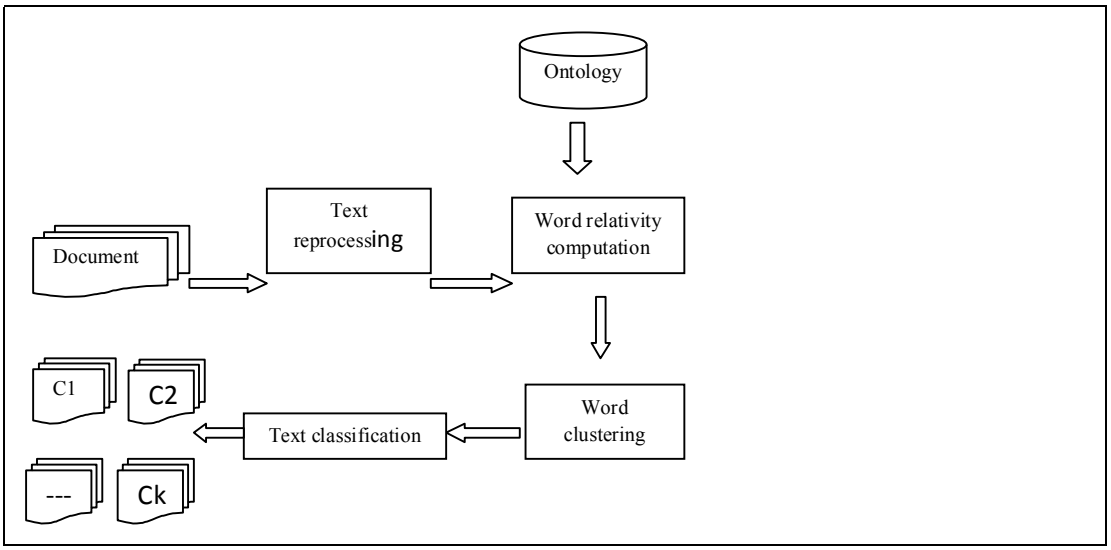

Figure 1: Word relativity-based clustering method

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.

Clustering technique is used for text categorization. Using this technique documents are classified into groups such that documents within any one group are closely related and documents in different groups are not closely related. The terms along with their weights are used for creating these groups. Each group or cluster is represented by a list of terms and those terms will appear in most of the documents within the group. SAS Text Miner uses Expectation-Maximization algorithm for clustering. For all the four decades we first used 20 for the maximum number of clusters property and subsequently modified this property based on clarity of clustering results (Shaik, Garla, Chakraborty, 2012).

Categorization

The process of recognizing, differentiating and understanding the ideas and objects to group them into categories, for specific purpose is Categorization. A category clarifies a relationship between the subjects and objects of knowledge. Categorization is essential in language, prediction, inference, decision making and in all kinds of environmental interaction.

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.

Categorization (Gupta, Lehal, 2009) 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.

Categorization is Relativity Analytics supervised mode of document classification. Whereas Clustering can be a fairly automated, hands-off process, Categorization always requires thorough and up-front preparatory work.

Conclusion

In this paper we have described the different text mining techniques such as Information Extraction, Information retrieval, Natural Language processing, Categorization, Query processing and Clustering.

In this short survey, the concept of text mining has been introduced and several techniques available have been presented. There are many prospective research areas in the field of Text Mining, which includes finding better in-between forms for representing the outputs of information extraction. In different languages text mining is a major problem, since these tools should be able to work with many languages and multilingual documents. Combining a domain knowledge base with a text mining engine would improve its efficiency, especially in the information retrieval and information extraction phases.

References

Ghosh S, Roy S, and Bandyopadhyay S K, (2012), A tutorial review on Text Mining Algorithms, *International Journal of Advanced Research in Computer and Communication Engineering,*1(4).

Gupta V, Lehal G S, (2009), A Survey of Text Mining Techniques and Applications, *Journal Of Emerging Technologies In Web Intelligence,* 1 (1).

Ramanathan V, Meyyappan T, (2013), Survey of Text Mining, *International Conference on Technology and Business Management*.

Sagayam R, Srinivasan S, and Roshni S, (2012), A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques,*International Journal Of Computational Engineering Research*, 2(5).

Shaik Z, Garla S, and Chakraborty G, (2012) SAS® Since 1976: An Application of Text Mining to Reveal Trends.

Winter School, Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets. Text Mining.

Witten, Ian H. *Text mining*.