A Survey Paper on Text Mining Techniques

C.Uma[1],S.Krithika[2],C.Kalaivani[3]

Assistant Professor, Department of CT&IT, Kongu Arts and Science College, Erode, India[1]
Assistant Professor, Department of Computer Science (PG), Kongu Arts and Science College, Erode, India [2]
Assistant Professor, Department of CT&IT, Kongu Arts and Science College, Erode, India [3]

***ABSTRACT** - Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. 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. We also discuss future challenges of this area using different techniques, particularly rough set based text mining techniques, improvements and research directions in this paper.*

***Keywords:** Data mining, Text mining, Rough sets, Classification, Summarization, and Text categorization, Clustering, Information Extraction, Information Retrieval.*

I. INTRODUCTION

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

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

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.

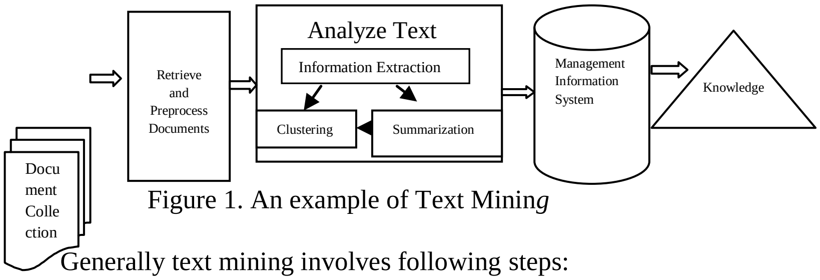


Figure 1. An example of Text Mining

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.

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

*(1) **Information Extraction***

Natural language text documents contain information that cannot be used for mining. As documents are considered as a bag of words, they can be represented by vector model which then can be exercised as an input to the above-defined techniques such as classifications, clustering but this is not used for this method. In Information extraction, the documents are first converted into the structured databases on which data mining techniques can be applied to extract knowledge or interesting patterns.

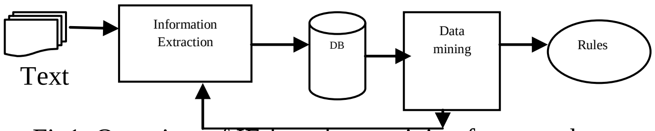


Fig1: Overview of IE-based text mining framework

In a relational database, data is stored in tables within rows and columns. A structured query on the database can help you retrieve the information required if the names of tables and columns are known. However, in the case of unstructured data, it is not easy to extract specific portions of information from the text because there is no fixed reference to identify the location of the data. Unstructured data can contain small fragments of information that might be of specific interest, based on the context of information and the purpose of analysis. Information extraction can be considered the process of extracting those fragments of data such as the names of people, organizations, places, addresses, dates, times, etc., from documents. Information extraction might yield different results depending on the purpose of the process and the elements of the textual data. Elements of the textual data within the documents play a key role in defining the scope of information extraction.

These elements are tokens, terms, and separators. A document consists of a set of tokens. A token can be considered a series of characters without any separators. A separator can be a special character, such as a blank space or a punctuation mark. A term can be a defined as a token with specific semantic purpose in a given language.

There are several types of information extraction that can be performed on textual data.

- ✓ Token extraction
- ✓ Term extraction or term parsing
- ✓ Concept extraction
- ✓ Entity extraction
- ✓ Atomic fact extraction
- ✓ Complex fact extraction

Concept extraction involves identifying nouns and noun phrases. Entity extraction can be defined as the process of associating Nouns with entities. For example, although the word “white” is a noun in English and represents a color, the occurrence of “Mr. White” in a document can be identified as a person, not a color. Similarly, the phrase “White House” can be attributed to a specific location (the official residence and principal workplace of the president of the United States), rather than as a description of the color of paint used for the exterior of a house. Atomic fact extraction is the process of retrieving fact –based Information based on the association of nouns with verbs in the content.

(2) 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.

Information retrieval, commonly known as IR, is the study of searching and retrieving a subset of documents from a universe of document collections in response to a search query. The documents are often unstructured in nature and contain vast amounts of textual data. The documents retrieved should be relevant to the information needs of the user who performed these arch queries. Several applications of the IR process have evolved in the past decade. One of them most ubiquitously known is searching for information on the World Wide Web. There are many search engines such as Google, Bing, and Yahoo facilitating this process using a variety of advanced methods. Most of the online digital libraries enable its users to search through their catalogs based on IR techniques. Many organizations enhance their websites with search capabilities to find documents, articles, and files of interest using keywords in the search queries. For example, the United States Patent and Trademark Office provide several ways of searching its database of patents and trademarks that it has made available to the public. In general, an IR system's efficiency lies n its ability to match a user's query with the most relevant documents in a corpus. To make the IR process more efficient, documents are required to be organized, indexed, and tagged with metadata based on the original content of the documents. SAS Crawler is capable of pulling information from a wide variety of data sources. Documents are then processed by parsers to create various fields such as title, ID, URL, etc., which form the metadata of the documents.

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}. 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.

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

Recall:

This is the percentage of documents that are relevant to the query and were, in fact, retrieved.

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

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.

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

(3) Natural Language processing

It deals with automatic processing and analysis of unstructured text information. It aims at processing the words found in text. In this two main fields are considered:

(i) Natural Language Generation (NLG): NLG uses linguistic representation of text, so that generated text is grammatically correct and fluent [3]. Most of these systems can include syntactic realizes ensuring that grammatical rules are efficiently followed or not. One example of NLG application is machine translation system [8].

(ii) Natural Language Understanding (NLU): NLU is a structure that finds the meaningful representation, by checking the discussion to the domain of computational language [5]. It contains at least one of these constituents: morphological or lexical analysis, tokenization, semantic analysis and syntactic analysis. Tokenization divides a sentence into list of tokens, which represents a special symbol or word. In morphological or lexical analysis each word is tagged with its part of speech, it becomes complex when a word contains more than one part of speech [4]. Syntactic analysis assigns parse tree to given natural language sentence, determining broking of sentence into phrases, sub phrases to actual word. In Semantic Analysis syntactic structure of sentence is translated into semantic representation, which allows system to perform appropriate task in its application domain [4][9].

Semantic interpretation contains two steps:

a) Context Independent Interpretation: that concerns with the meaning of words and combining these meanings into sentences to find meaning of sentences.

b) Context Interpretation: it concerns with effect of context on interpretation of sentences [9]. Context includes situation of usage of sentence, preceding sentences etc.

(4) 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 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.

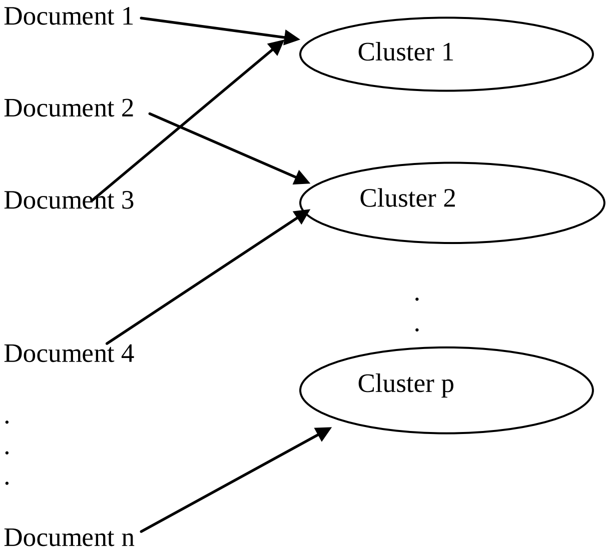
The first is the synonymy problem: two words with identical or similar meanings may have very different surface forms. For example, a user's query may use the word "automobile," but a relevant document may use "vehicle" instead of "automobile."

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

(5) Clustering

Cluster analysis is a popular technique used by data analysts in numerous business applications. Clustering partitions records in a dataset into groups so that the subjects within a group are similar and the subjects between the groups are dissimilar. The Goal of cluster analysis is to derive clusters that have value with respect to the problem being addressed, but this goal is not always achieved. As a result, there are many competing clustering algorithms. The analyst often compares the quality of derived clusters, and then selects the method that produces the most useful groups. The clustering process arranges documents into non overlapping groups. Each document can fall into more than one topic area after classification. This is the key difference between clustering and the general text classification processes, although clustering provides a solution to text classification when groups must be mutually exclusive, as in the classified ads example.

In the context of text mining, clustering divides the document collection into mutually exclusive groups based on the presence of similar themes. In most business applications involving large amounts of textual data, it is often difficult to profile each cluster by manually reading and considering all of the text in a cluster. Instead, the theme of a cluster is identified using a set of descriptive terms that each cluster contains. This vector of terms represents the weights measuring how the document fits into each cluster. Themes help in better understanding the customer, concepts, or events. The number of clusters that are identified can be controlled by the analyst. The algorithm can generate clusters based on the relative positioning of documents in the vector space. The cluster configuration is altered by a start and stop list.



[Fig 2] Text Clustering Process Assigning Each Document to Only One Cluster

For example, consider the comments made by different patients about the best thing that they liked about the hospital that they visited.

1. Friendliness of the doctor and staff.
2. Service at the eye clinic was fast.
3. The doctor and other people were very, very friendly.
4. Waiting time has been excellent and staff has been very helpful.
5. The way the treatment was done.
6. No hassles in scheduling an appointment.
7. Speed of the service.
8. The way I was treated and my results.
9. No waiting time, results were returned fast and great treatment.

The clustering results from text mining the comments come out similar to the ones shown in Table 1.Each cluster can be described by a set of terms, which reveal, to a certain extent, the theme of the cluster. This type of analysis helps businesses Understand the collection as a whole, and it can assist in correctly classifying customers based on common topics in customer Complaints or responses.

Cluster No.	**Comment**	**Key Words**
1	1,3,4	doctor, staff, friendly, helpful
2	5, 6, 8	treatment, results, time, schedule
3	2, 7	service, clinic, fast

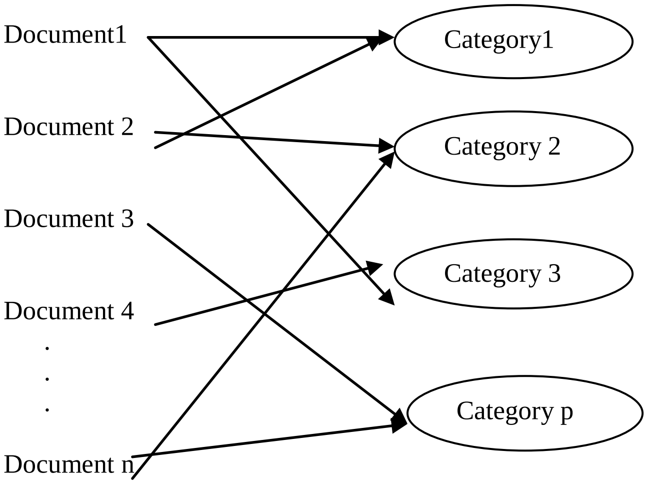
Table 1 Clustering Results from Text Mining

(6) Classification

Document classification is the process of finding commonalities in the documents in a corpus and grouping them into predetermined labels (supervised learning) based on the topical themes exhibited by the documents. Similar to the IR process, document classification (or text categorization) is an important aspect of text analytics and has numerous applications. Some of the common applications of document classification are e- mail forwarding and spam detection, call center routing, and news articles categorization. It is not necessary that documents be assigned to mutually exclusive categories.

Any restrictive approach to do so might prove to be an inefficient way of representing the information. In reality, a document can exhibit multiple themes, and it might not be possible to restrict them to only one category.

In cases where a document should be restricted to only one category, text clustering is usually a better approach instead of extracting text topics. For example, an analyst could gain an understanding of a collection of classified ads when the clustering algorithm reveals the collection actually consists of categories such as Car Sales, Real Estate, and Employment Opportunities.



[Fig 3] Text Categorization Involving Multiple Categories per Document

II 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.

REFERENCE

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

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

[3] Shaidah Jusoh and Hejab M. Alfawareh, “Techniques, Applications and Challenging Issue in Text Mining”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012, ISSN (Online): 1694-0814.

[4] Fred Popowich, “Using Text Mining and Natural Language Processing for Health Care Claims Processing”

[5] Hejab M. Alfawareh, Shaidah Jusoh, “Resolving Ambiguous Entity through Context Knowledge and Fuzzy Approach”, International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397 Vol. 3 No. 1 Jan 2011

[6] D. Cutting, D. Karger, J. Pedersen, J. Tukey. Scatter/Gather: ACluster-based Approach to Browsing Large Document Collections.ACM SIGIR Conference, 1992.

[7] L. Baker, A. McCallum. Distributional Clustering of Words for Text Classification,ACM SIGIR Conference , 1998.

[8] R. Bekkerman, R. El-Yaniv, Y. Winter, N. Tishby. On Feature Dis-tributional Clustering for Text Categorization.ACM SIGIR Con-ference, 2001.

[9] K. Nigam, A. McCallum, S. Thrun, T. Mitchell. Learning to classify text from labeled and unlabeled documents.AAAI Conference,1998.