RESEARCH ARTICLE OPEN ACCESS

A Review on Text Mining Techniques

S.Sathya [1], Dr.N.Rajendran [2]
Research Scholar [1]
Bharathiar University, Coimbatore
Principal [2]
Vivekanandha Arts and Science College for Women, Sankari
Tamil Nadu - India

ABSTRACT

Text mining refers commonly to the method of extracting interesting information and knowledge from unstructured text. Text Mining has become an important research area to extract useful and interesting information from this large amount of textual data. Text mining is the way of discovering knowledge from previously unknown text data, by automatically extracting information from different resources. In this article, we discuss text mining techniques like Data Mining, Information Retrieval, Information extraction, natural language processing, summarization, categorization, topic discovery, clustering, and concept linkage and information visualization. In addition, we discussed the applications of text mining.

Keywords:- Text mining, information retrieval, extraction, Text Categorization, summarization, Clustering

I.INTRODUCTION

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 preprocessing steps in order to extract data from texts. To the extracted data then data mining algorithms can be applied [1] [2]. Text mining, sometimes alternately 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 divining of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Regarded by many as the next wave of knowledge discovery, text mining has very high commercial values.

II. TEXT MINING TECHNIQUES

Text Mining (TM) refers some informational content included in any of the items such as: newspaper; articles; books; reports; stories; manuals; blogs; email, and articles in the www. The quantum of text of the present day is pretty vast with ever-growing incremental power [3]. The prime aim of the text mining is to identify the useful information without duplication from various documents with synonymous understanding. TM is an empirical tool that has a capacity of identifying new information that is not apparent from a document collection. Figure 1 depicts the TM process that uses Information retrieval and Natural Language Processing to mine large dataset and infer the knowledge available in the dataset. The Process of TM includes searching, extracting, categorization where the themes are readable and the meaning is obvious. Typically text mining tasks include text categorization, text clustering, Information extraction, information retrieval, sentiment analysis, document summarization, and entity relation modeling. TM, also known as Knowledge Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT) refers generally to the process of extracting information and knowledge from unstructured text[3][4][5].

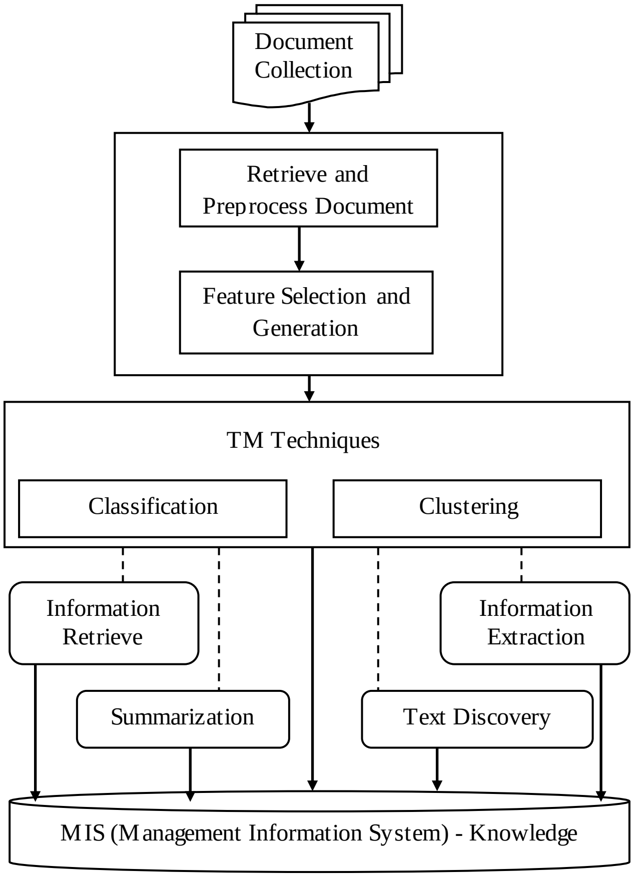


Figure 1 Text Mining process

TM starts with a collection of documents; which would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. Three text analysis techniques are shown in the example, but many other combinations of techniques could be used depending on the goals of the organization. The resulting information can be placed in a management information system, yielding an abundant amount of knowledge for the user of that system. The following figure explores the detail processing methods in Text Mining.

a) Data Mining:

DM is the process of identifying patterns in large sets of data. The aim is to uncover previously unknown, useful knowledge. When used in text mining, DM is applied to the facts generated by the information extraction phase. We put the results of our DM process into another database that can be queried by the end-user via a suitable graphical interface. The data generated by such queries can also be represented visually.

Data mining can be loosely described as looking for patterns in data. It can be more fully characterized as the extraction of hidden, previously unknown, and useful information from data. Data mining tools can predict behaviors and future trends, allowing businesses to make positive, knowledge based decisions. Data mining tools can answer business questions that have traditionally been too time consuming to resolve. They search databases for hidden and unknown patterns, finding critical information that experts may miss because it lies outside their expectations. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use [7].

b) Information Retrieval:

IR systems finding the documents in a collection which match a user's query.[8] The most well known IR systems are search engines such as Google, which identify those documents on the World Wide Web that are relevant to a set of given words. 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 [9]. Although, information retrieval is a relatively old research area where first attempts for automatic indexing where made in 1975 [10], it gained increased attention with the rise of the World Wide Web and the need for sophisticated search engines. IR systems are often used in libraries, where the documents are typically not the books themselves but digital records containing information about the books. IR systems allow us to narrow down the set of documents that are relevant to a particular problem. As text mining involves applying very computationally-intensive algorithms to large document collections, IR can speed up the analysis considerably by reducing the number of documents for analysis. For example, if we are interested in mining information only about protein interactions, we might restrict our analysis to documents that contain the name of a protein, or some form of the verb 'to interact' or one of its synonyms.

Models of Information Retrievals:

Boolean Model: A document is represented by a set of key terms: chosen from a fixed set of key terms, or, possibly automatically, from the documents themselves.

Vector Model: In a V*ector-model* IR system containing *n* key terms, an *n* dimensional space is defined such that each axis is associated with a different key term.

Probabilistic Model: The basis for the probabilistic model is the probability ranking principle best possible retrieval results are achieved when documents are shown in the order of their probable relevance to the Query.

Connectionist Model: Neural networks create a form of connectionism this is also applicable in IR. IR purposes each key term can be associated with an input neuron and each document with an output neuron. A query is presented to the network by activating the neurons which are associated with the desired key terms.

c) Natural Language Processing:

NLP is one of the oldest and most difficult problems in the field of artificial intelligence. Text and data mining approach has been used in Natural language processing to overcome the difficulties with the codes, keywords and search techniques involved in knowledge or pattern discovery. The output of this mining process helps analyst to discover the trends and new occurrences in the data available. Many TM algorithms have been developed to maintain the text sources of bilingual corpus or multi lingual corpus. The general goal of NLP is to achieve a better understanding of natural language by use of computers [11]. Others include also the employment of simple and durable techniques for the fast processing of text, as they are presented e.g. in [12]. The range of the assigned techniques reaches from the simple manipulation of strings to the automatic processing of natural language inquiries.

In addition, linguistic analysis techniques are used among other things for the processing of text. It is the analysis of human language so that computers can understand natural languages as humans do. Although this goal is still some way off, NLP 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 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.

Natural Language processing, Text mining and Machine learning methods can be applied for mining of interesting data available online for example gene ontology or protein function mapping using certain tools. For example Rapier and Tagging process is done through for certain Biomedical corpus available. Rapier is a machine learning tool that learns information extraction rules from a set of documents and associated templates. This form of representation is grammar rule induction. The Rapier works on tagged documents directly instead of templates. We ran this instance of Rapier on our manually tagged training corpus to produce a set of grammar rules [13].

d) Information Extraction:

IE is the process of mechanically obtaining structured data from an unstructured natural language document. Often 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 [14]. A starting point for computers to examine unstructured text is to use information extraction. Information extraction software identifies key phrases and relationships within text. The software infers the relations between all the identified people, places, and time to deliver the user with significant information. This technology can be very helpful when dealing with large volumes of text. Traditional data mining assumes that the information to be “mined” is previously in the form of a relational database. Unfortunately, for many applications, electronic information is only obtainable in the form of free natural language documents rather than structured databases. Since IE addresses the difficulty of transforming a corpus of textual documents into a extra structured database, the database constructed by an IE module can be provided to the KDD module for advance mining of knowledge as illustrated in Figure 2.

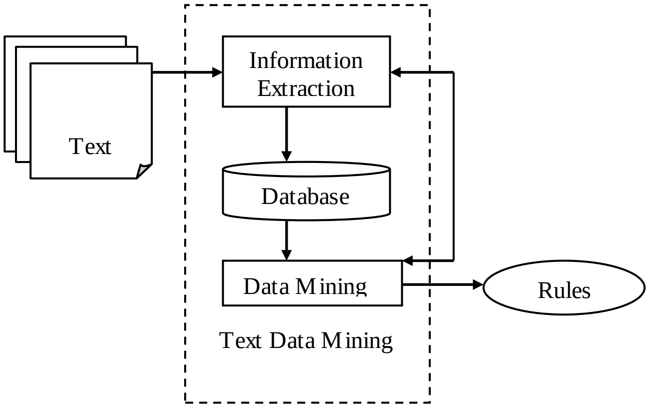


Figure 2: Information Extraction Modules

e) Topic Tracking:

A topic tracking system mechanism by custody of user profiles and, based on the documents the user views, guess other documents of interest to the user. Yahoo offers a free topic tracking tool (www.alerts.yahoo.com) that allows users to choose keywords and notifies them when news relating to those topics becomes available. Topic tracking technology does have limitations, however. For example, if a user sets up an alert for "text mining", user will receive several news stories on mining for minerals, and very few that are actually on text mining. Some of the better text mining tools let users select particular categories of interest or the software automatically can even infer the user's interests based on his/her reading history and click-through information. There are many areas where topic tracking can be applied in industry. It can be used to alert companies anytime a competitor is in the news. This allows them to keep up with competitive products or changes in the market. Similarly, businesses might want to track news on their own company and products. It could also be used in the medical industry by doctors and other people looking for new treatments for illnesses and who wish to keep up on the latest advancements. Individuals in the field of education could also use topic tracking to be sure they have the latest references for research in their area of interest [15].

Keywords are a set of significant words in an article that gives high-level description of its contents to readers. Identifying keywords from a large amount of on-line news data is very useful in that it can produce a short summary of news articles. As on-line text documents rapidly increase in size with the growth of WWW, keyword extraction [16] has become a basis of several text mining applications such as search engine, text categorization, summarization, and topic detection. Manual keyword extraction is an extremely difficult and time consuming task; in fact, it is almost impossible to extract keywords manually in case of news articles published in a single day due to their volume. For a rapid use of keywords, we need to establish an automated process that extracts keywords from news articles.

The architecture of keyword extraction system is presented in figure 3. HTML news pages are gathered from a Internet portal site. And candidate keywords are extracted throw keyword extraction module. And finally keywords are extracted by cross-domain comparison module. Keyword extraction module is described in detail. We make tables for 'document', 'dictionary', 'term occur fact' and 'TFIDF weight' in relational database. At first the downloaded news documents are stored in 'Document' table and nouns are extracted from the documents in 'Document table.

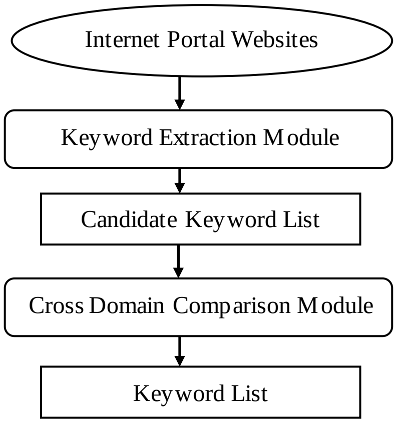


Figure 3: The architecture of keyword extraction system

Then the facts which words are appeared in documents are updated to 'Term occur fact' table. Next, TF-IDF weights for each word are calculated using 'Term occur fact' table and the result are updated to 'TF-IDF weight' table. Finally, using 'TF-IDF weight' table, 'Candidate keyword list' for each news domain with words is ranked high. Keyword extraction module is given in figure 4.

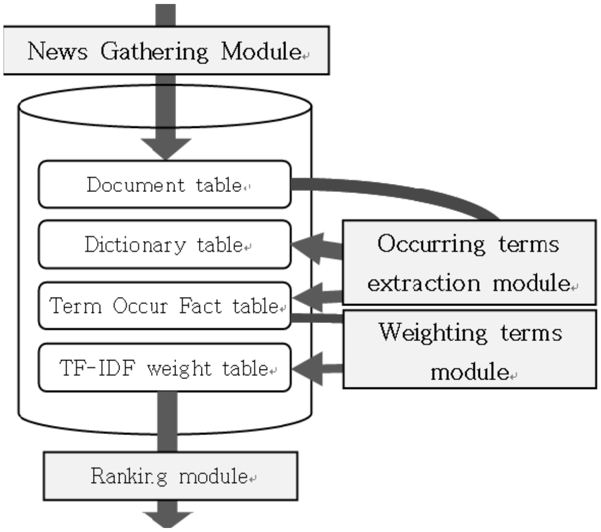


Figure 4: Keyword extraction module

Lexical chaining [17] is a method of grouping lexically related terms into so called lexical chains. Topic tracking involves tracking a given news event in a stream of news stories i.e. finding all the subsequent stories in the news stream. In multi vector topic tracking system proper names, locations and normal terms are extracted into distinct sub vectors of document representation. Measuring the similarity of two documents is conducted by comparing two sub-vectors at a time. Numbers of features that affect the performance of topic tracking system are analyzed. First choice is to choose one characteristic, such as the choice of words, words or phrases such as string as a feature in this term to make features as an example [18].

f) Text summarization:

Text summarization is immensely helpful for trying to figure out whether or not a lengthy document meets the user's needs and is worth reading for further information. With large texts, text summarization software processes and summarizes the document in the time it would take the user to read the first paragraph. The key to summarization is to reduce the length and detail of a document while retaining its main points and overall meaning. The challenge is that, although computers are able to identify people, places, and time, it is still difficult to teach software to analyze semantics and to interpret meaning. Generally, when humans summarize text, we read the entire selection to develop a full understanding, and then write a summary highlighting its main points. Since computers do not yet have the language capabilities of humans, alternative methods must be considered. One of the strategies most widely used by text summarization tools, sentence extraction, extracts important sentences from an article by statistically weighting the sentences. Further heuristics such as position information are also used for summarization [19].

For example, summarization tools may extract the sentences which follow the key phrase "in conclusion", after which typically lie the main points of the document. Summarization tools may also search for headings and other markers of subtopics in order to identify the key points of a document. Microsoft Word's AutoSummarize function is a simple example of text summarization. Many text summarization tools allow the user to choose the percentage of the total text they want extracted as a summary. Summarization can work with topic tracking tools or categorization tools in order to summarize the documents that are retrieved on a particular topic. If organizations, medical personnel, or other researchers were given hundreds of documents that addressed their topic of interest, then summarization tools could be used to reduce the time spent sorting through the material. Individuals would be able to more quickly assess the relevance of the information to the topic they are interested in.

The methods of summarization can be classified, in terms of the level in the linguistic space, in two broad groups: [20]

(a) Shallow approaches, which are restricted to the syntactic level of representation and try to extract salient parts of the text in a convenient way.
(b) Deeper approaches, which assume semantics level of representation of the original text and involve linguistic processing at some level.

In the approaches the aim of the preprocessing step is to reduce the dimensionality of the representation space, and it normally includes:

(i) Stop-word elimination : common words with no semantics and which do not aggregate relevant information to the task (e.g., "the", "a") are eliminated

(ii) Case folding: consists of converting all the characters to the same kind of letter case - either upper case or lower case

(iii) Stemming: syntactically-similar words, such as plurals, verbal variations, etc. are considered similar; the purpose of this procedure is to obtain the stem or radix of each word, which emphasize its semantics.

Text summarization involves various methods that employ text categorization, such as neural networks, decision trees, semantic graphs, regression models, fuzzy logic and swarm intelligence. However, all of these methods have a common problem, that is, the quality of the development of classifiers is variable and highly dependent on the type of text being summarized [21].

g) Text Categorization:

Categorization engages 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 delight the document as a "bag of words."Rather, categorization only calculates words that emerge and, from the counts, identifies the main topics that the document covers. Categorization often relies on a vocabulary for which topics are predefined, and relationships are recognized by looking for broad terms, narrower terms, synonyms, and related terms. Categorization utensils normally have a technique for grade the documents in order of which documents have the most content on a specific topic. As with summarization, categorization can be used with topic tracking to further specify the relevance of a document to a person seeking information on a topic. The documents returned from topic tracking could be ranked by content weights so that individuals could give priority to the most relevant documents first. Categorization 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 [22].

Text categorization is a kind of "supervised" learning where the categories are known in advance and firm in progress for each training document. Then, its key projected utilize was for indexing scientific literature by means of controlled words [23]. Categorization is the assignment of normal language documents to predefined set of topics according to their content. It is a collection of text documents, the process of finding the accurate topic or topics for each document. Nowadays automated text categorization is applied in a variety of contexts from the classical automatic or semiautomatic indexing of texts to personalized commercials delivery, spam filtering, and categorization of Web page under hierarchical catalogues, automatic metadata generation, and detection of text genre, topic tracking and many others [24].

Using supervised learning algorithms [25], the objective is to learn classifiers from known examples (labeled documents) and perform the classification automatically on unknown examples (unlabeled documents). Figure.8 shows the overall flow diagram of the text categorization task. Consider a set of labeled documents from a source $D = [d1,d2,….dn]$ belonging to a set of classes $C = [c1,c2,…,cp]$. The text categorization task is to train the classifier using these documents, and assign categories to new documents. In the training phase, the n documents are arranged in p separate folders, where each folder corresponds to one class. In the next step, the training data set is prepared via a feature selection process.

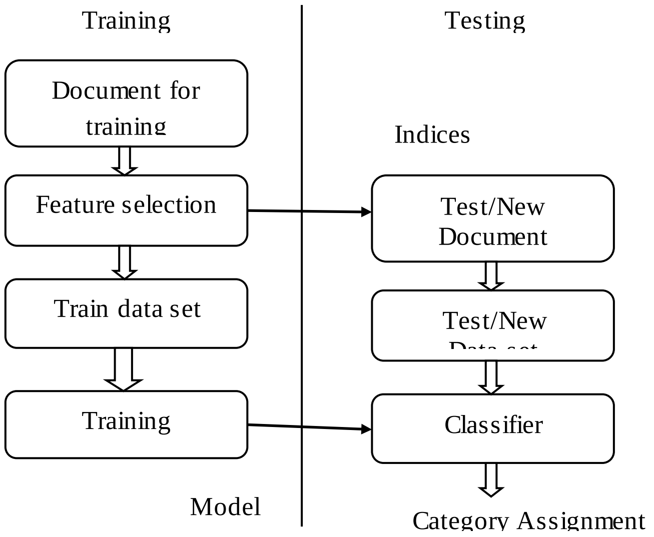
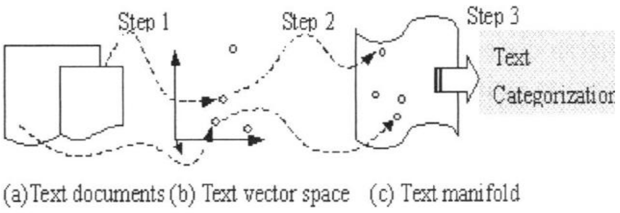


Figure 5: Flow Diagram of Text Categorization.

Text data typically consists of strings of characters, which are transformed into a representation suitable for learning. It is observed from previous research that words work well as features for many text categorization tasks. In the feature space representation, the sequences of characters of text documents are represented as sequence of words. Feature selection involves tokenizing the text, indexing and feature space reduction. Text can be tokenized using term frequency (TF), inverse document frequency (IDF), term frequency inverse document frequency (TFIDF) or using binary representation. Using these representations the global feature space is determined from entire training document collection. In text manifold [26] categorization method, the text documents are treated as vectors in an n-dimensional space, where every dimension corresponds to a term. Then the metrics such as the cosine of the angle between two documents can be defined. However this space may be intrinsically located on the low dimensional manifold. The metric therefore should be defined according to the properties of manifold so as to improve the text categorization furthermore. The whole process is illustrated as Figure 6.



(a)Text documents (b) Text vector space (c) Text manifold

Figure 6: Text Categorization Process

h) Clustering:

Clustering is a technique used to group similar documents, but it differs from categorization in that documents are clustered on the fly instead of through the use of predefined topics. Its aim is to find intrinsic structures in information, and arrange them into significant subgroups for further study and analysis. It is an unsupervised process through which objects are classified into groups called clusters. The problem is to group the given unlabeled collection into meaningful clusters without any prior information. Any labels associated with objects are obtained solely from the data. For example, document clustering assists in retrieval by creating links between related documents, which in turn allows related documents to be retrieved once one of the documents has been deemed relevant to a query. Document clustering has been studied intensively because of its wide application in areas such as Web Mining, Search Engine and Information Retrieval. Document clustering is the automatic organization of documents into clusters or groups, so that, documents within a cluster have high similarity in comparison to one another, but are very dissimilar to documents in other clusters [27]. Clustering is useful in many application areas such as biology, data mining, pattern recognition, document retrieval, image segmentation, pattern classification, security, business intelligence and Web search. Cluster analysis can be used as a standalone text mining tool to achieve data distribution, or as a pre-processing step for other text mining algorithms operating on the detected clusters.

i) Concept Linkage:

Concept linkage tools attach related documents by identifying their commonly-shared idea and help users find information that they perhaps wouldn't have establish using conventional searching methods. It promotes browsing for information rather than searching for it. This mechanism browses documents instead of search. It offers the facility to link related documents. This technique is useful in many areas such as medical field to find documents related to diseases and treatment so that it helps to doctor very efficiently. Government also uses concept linkage for criminal records with previous records for getting idea about criminal and its relationship [28]. Concept linkage is a valuable idea in text mining, especially in the biomedical fields where so much study has been done that it is impossible for researchers to read all the material and make organizations to other research. Ideally, concept linking software can identify links between diseases and treatments when humans cannot. For example, a text mining software solution may easily identify a link between topics X and Y, and Y and Z, which are well-known relations. But the text mining tool could also detect a potential link between X and Z, something that a human researcher has not come across yet because of the large volume of information s/he would have to sort through to make the connection [29].

j) Information Visualization:

Visual text mining is a simple searching for extracting the patterns. Visual text mining, or information visualization, puts large textual sources in a visual hierarchy or map and provides browsing capabilities and simple searching. Tools of Visual text mining helps to user can interact with the document map by zooming, scaling, and creating sub-maps. Information visualization is useful when a user needs to narrow down a broad range of documents and explore related topics. The government can use information visualization to identify terrorist networks or to find information about crimes that may have been previously thought unconnected. It could provide them, with a map of possible relationships between suspicious activities so that they can investigate connections that they would not have come up with on their own [30].

The goal of information visualization, the construction may be conducted into three steps:

- **Data preparation:** i.e. determine and acquire original data of visualization and form original data space.
- **Data analysis and extraction:** i.e. analyze and extract visualization data needed from original data and form visualization data space.
- **Visualization mapping:** i.e. employ certain mapping algorithm to map visualization data space to visualization target.

InforVisModel tools divide the construction into five steps:

- Information collection: to collect information resources needed from databases or WWW.
- Information indexing: to index collected information resources to form original data sources.
- Information retrieval: to query information lists in conformity to result from original data sources according to the need of retrieval.
- Generation of visualization data: to transform data in the retrieved results into visualization data.
- Display of visualization interface: to map visualization data to visualization target and display them on visualization interface [31].

k) Question Answering:

Another application area of natural language processing is natural language queries, or question answering (Q&A), which deals with how to find the best answer to a given question. Many websites that are equipped with question answering technology, allow end users to “ask” the computer a question and be given an answer. Q&A can utilize multiple text mining techniques. For example, it can use information extraction to extract entities such as people, places, events; or question categorization to assign questions into known types (who, where, when, how, etc.). In addition to web applications, companies can use Q&A techniques internally for employees who are searching for answers to common questions. The education and medical areas may also find uses for Q&A in areas where there are frequently asked questions that people wish to search.

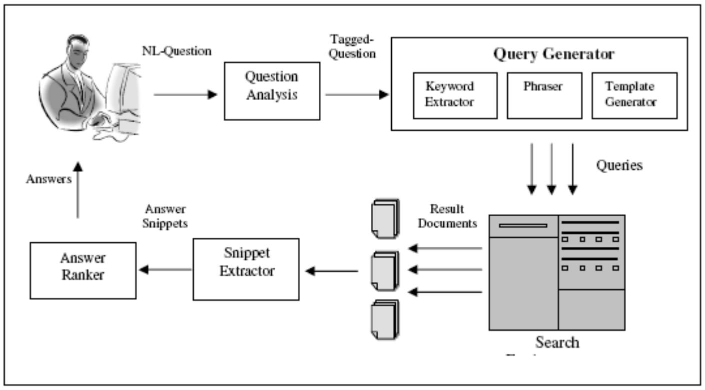


Figure 7: Architecture of Question answering system

Figure 7 shows the architecture of question answering system [32]. The system takes in a natural language (NL) question in English from the user. This question is then passed to a Part-of-Speech (POS) tagger which parses the question and identifies POS of every word involved in the question. This tagged question is then used by the query generators which generate different types of queries, which can be passed to a search engine. These queries are then executed by a search engine in parallel. The search engine provides the documents which are likely to have the answers we are looking for. These documents are checked for this by the answer extractor. Snippet Extractor extracts snippets which contain the query phrases/words from the documents. These snippets are passed to the ranker which sorts them according to the ranking algorithm.

III. CONCLUSION

Text mining is a young interdisciplinary playing field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. Text Mining can be defined as a technique which is used to extract interesting information or knowledge from the text documents which are usually in the unstructured form. Various techniques for efficiently performing text mining are discussed in this paper. So in this paper, our focus is basically on how text is to be mined. We have also discussed process of text mining and its applications.

REFERENCES

[1] 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.

[2] R. Gaizauskas. An information extraction perspective on text mining: Tasks, technologies and prototype applications. http://www.itri.bton.ac.uk/2003.

[3] Shu-Sheng Liaw, Hsiu-Mei Huang, "Information Retrieval from the World Wide Web: a User-focused Approach based on Individual Experience with Search Engines". *Computers in Human Behavior*, **22** (2006).

[4] Setu Madhavi Namburu, Haiying Tu, Jianhui Luo and Krishna R. Pattipati (2005), "Experiments on Supervised Learning Algorithms for Text Categorization", *International Conference, IEEE Computer Society*, 1-8.

[5] Vishal Gupta, Gurpreet S. Lehal. "A Survey of Text Mining Techniques and Applications". *Journal of Emerging Technologies in Web Intelligence*, **1**, No. 1, August 2009".

[6] Zhou Ning, Wu Jiaxin, Wang Bing and Zhang Shaolong (2008), "A Visualization Model for Information Resources Management", *12th International Conference Information Visualisation*, China**,** *IEEE*, 57- 62.

[7] Mrs. Sayantani Ghosh, Mr. Sudipta Roy, and Prof. Samir K. Bandyopadhyay. ―A tutorial review on Text Mining Algorithms‖, in International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 4, 2012.

[8] M. Hearst. Untangling text data mining. In *Proc. of ACL'99 the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

[9] K. Sparck-Jones and P. Willett, editors. *Readings in Information Retrieval*. Morgan Kaufmann, 1997.

[10] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975. (see also TR74-218, Cornell University, NY, USA).

[11] Y. Kodratoff. Knowledge discovery in texts: A definition and applications. *Lecture Notes in Computer Science*, 1609:16–29, 1999.

[12] S. P. Abney. Parsing by chunks. In R. C. Berwick, S. P. Abney, and C. Tenny, editors, *Principle-Based Parsing: Computation and Psycholinguistics*, pages 257–278. Kluwer Academic Publishers, Boston, 1991.

[13] Jiexun Li, Harry Jiannan Wang, Zhu Zhang and J. Leon Zhao "A Policy-based Process Mining Framework: Mining Business Policy Texts for Discovering Process Models" *Information System s and E-Business Management*, 2010.

[14] Y. Wilks. Information extraction as a core language technology. In M-T. Pazienza, editor, *Information Extraction*. Springer, Berlin, 1997.

[15] J. Allan, editor. Topic Detection and Tracking. Kluwer Academic Publishers, Norwell, MA, 2002.

[16] Sungjick Lee and Han-joon Kim (2008), "News Keyword Extraction for Topic Tracking", Fourth International Conference on Networked Computing

and Advanced Information Management, IEEE, Koria,554-559.

[17] Joe Carthy and Michael Sherwood-Smith (2002), “Lexical chanins for topic tracking”, International Conference, IEEE SMC WP1M1, Ireland.

[18] Wang Xiaowei, JiangLongbin, MaJialin and Jiangyan (2008), “Use of NER Information for Improved Topic Tracking”, Eighth International Conference on Intelligent Systems Design and Applications, IEEE computer society, Shenyang, 165-170.

[19] Vishal Gupta and Guruprit Lehal, “A Survey of Text Mining Techniques and Applications”, Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, August 2009.

[20] Farshad Kyoomarsi ,Hamid Khosravi ,Esfandiar Eslami ,Pooya Khosravyan Dehkordy and Asghar Tajoddin (2008), “Optimizing Text Summarization Based on Fuzzy Logic”, Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE computer

[21] society, 347-352.

[22] Fang Chen, Kesong Han and Guilin Chen (2008), “An approach to sentence selection based text summarization”,Proceedings of IEEE TENCON02, 489- 493.

[23] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In C. Nedellec and C. Rouveirol, editors, *European Conf. on Machine Learning (ECML)*, 1998.

[24] Hearst, M. A. (1997) Text data mining: Issues, techniques, and the relationship to information access. Presentation notes for UW/MS workshop on data mining, July 1997.

[25] JIAN-SUO XU (2007), “TCBPLK: A new method of text categorization ”, Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong,, IEEE , 3889-3892.

[26] Setu Madhavi Namburu, Haiying Tu, Jianhui Luo and Krishna R. Pattipati (2005), “Experiments on Supervised Learning Algorithms for Text Categorization”, International Conference , IEEE computer society, 1-8.

[27] Guihua Wen, Gan Chen, and Lijun Jiang (2006), “Performing Text Categorization on Manifold”, 2006 IEEE International Conference on Systems, Man, and Cybernetics, Taipei, Taiwan , IEEE , 3872-3877.

[28] Mr. Rahul Patel,Mr. Gaurav Sharma,"A survey on text mining techniques", International Journal Of Engineering And Computer Science ISSN:2319-7242, Vol 3 Issue 5, May 2014, pp.5621-5625

[29] Vishal gupta and Gurpreet S. Lehal , “A survey of text mining techniques and applications”, journal of emerging technologies in web intelligence, 2009,pp.60-76.

[30] Qing Cao, Wenjing Duan, Qiwei Gan, “Exploring determinants of voting for the “helpfulness” of online user reviews: A text mining approach’, 0167-9236/$ – see front matter © 2010 Elsevier B.V. All rights reserved. doi:10.1016/j.dss.2010.11.009

[31] N. Kanya and S. Geetha ,“Information Extraction: A Text Mining Approach”, IET-UK International Conference on Information and Comm. Technology in Electrical Sciences, IEEE(2007), Dr. M.G.R. University, Chennai, Tamil Nadu, India,1111- 1118

[32] Zhou Ning, Wu Jiaxin, Wang Bing and Zhang Shaolong (2008), “A Visualization Model for Information Resources Management”, 12th International Conference Information Visualisation, China, IEEE, 57- 62.

[33] Jignashu Parikh and M. Narasimha Murty (2002),“Adapting Question Answering Techniques to the Web”, Proceedings of the Language

Engineering Conference, India, IEEE computer society.

AUTHOR'S PROFILE

S.Sathya received her M.Phil(C.S) Degree from Bharathidasan University,Trichy in the year 2006. She has received her M.Sc.,(CS) Degree from Periyar University, Salem in the year 2003. She is working as Assistant Professor, Department of Computer Science, Vivekanandha College for Women, Namakkal, Tamilnadu, India. Her areas of interest include Database Management System, Data Mining and Networking.

Dr.N.Rajendran received his Ph.D Degree from Periyar University, Salem in the year 2011. He has received his M.Phil, Degree from Bharathiar University, Coimbatore in the year 2000. He has received his M.C.A Degree from Madras University, Chennai in the year 1990. He is working as Principal of Vivekanandha Arts and Science College for Women, Sankari, Salem , Tamilnadu, . He has 24 years of experience in academic field. He has published 15 International Journal papers and 13 papers in National and International Conferences. His areas of interest include Digital Image Processing and Networking.