ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Study of Text Content Mining for E-Commerce Web Sites

A.K. Shingarwade
Department of Computer Science, College of Management
& Comp. Sci., Yavatmal, (M.S.), India

Dr. P.N. Mulkalwar
Department of Computer Science, Amolakchand
Mahavidyalaya, Yavatmal (M.S.), India

Abstract: In this digital age, the traditional buying system of the consumers has affected. Now a large number of people, therefore have shifted from off line buying to online buying. Online buying helps in understanding user behavior like what are the likes and dislikes of the user, their buying behavior, and many such applications. There are multiple methods available and used from which the knowledge is mined from the data. In this paper we are going to particularly focus on text mining, the techniques studies and analyzed are based on text only. The techniques such as clustering, summarization and much more are studied and analyzed in the paper.

Keywords: Data Mining, text mining, E-commerce data, Summarization, Clustering.

I. **INTRODUCTION**:

Data collected from various sources is increasing day by day at exponential rate since all institutions and industries are storing data electronically. Data mining is one of the important research areas to scrape knowledge that can be the behavior of the customer, sales, losses which can be studied from various dimensions with the help of the data gathered [1]. There is a lot of data in different formats can be easily found in the sources. In this paper we are going to learn about text data, a lot of textual data in the form of digital libraries, repositories, and other textual information such as blogs, social media network, and e-mails [2] is present. When the volume of the data is large [3] it is difficult by the traditional data mining techniques to process the patterns from the data, since they are incapable of handling large textual data because of the delay and effort require to extract the patterns.

Textual mining is a procedure to separate intriguing and significant examples to investigate learning from textual information sources. Textual mining is a multi-disciplinary field in view of data recovery, information mining, machine learning, insights, and computational semantics [4]. Figure 1 demonstrates the Venn outline of textual mining and its communication with different fields. A few textual mining strategies like a rundown, classification, bunching and so forth, can be connected to concentrate learning. Textual mining manages characteristic dialect textual which is put away in the semi-organized and unstructured arrangement [5]. Textual mining methods are consistently connected in the industry, the scholarly community, web applications, the web, and different fields [6]. Application zones like web crawlers, client relationship administration framework,filter messages, item recommendation examination, misrepresentation recognition, and online networking investigation utilize textual digging for assessment mining, include extraction, opinion, prescient, and incline examination.

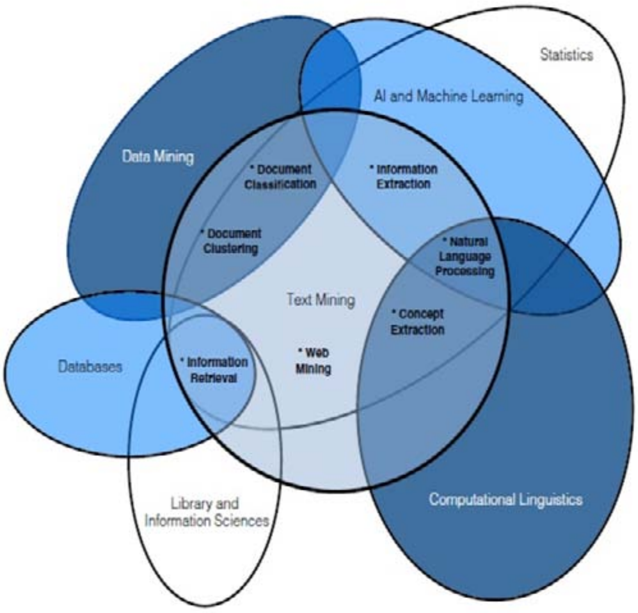


Figure 1: Venn diagram of text mining [1]

These advances have helped science, business, and analytical applications. In business, successful forecasts can be produced using anciently accessible information. The retail industry, money related, correspondence and showcasing associations which have solid customer center utilized information digging fundamentally for business expectations. These forecasts empower organizations to decide connections amongst inner and outside elements. Interior components may incorporate financial pointers, rivalry, and client socioeconomics. Information mining help clients coordinate the fate of their exercises by conveying exact and valuable information since choices here are made on sound database insight and not on impulses on feelings. In this paper section II deals with various techniques used for mining textual data following with their analysis and the conclusion concludes the paper.

II. WORKING OF TEXT MINING:

a. Traditional search techniques are used to search predefined keywords from the documents after which more accurate information like concepts, sentences, phrases, relationships etc. are extracted.

b. Text mining tool is used to extract the meaning of the text, identify, synthesize, extract and analyze relevant facts and relationships. Computational algorithms based on Natural language processing (NLP) is used by these tools which enable the computer to read and analyze textual information.

c. The text is mined in a systematic, comprehensive and reproducible way so as to capture the needed information repeatedly.

d. NLP based queries which are pre-written can be run in real time across millions of documents.
e. In order to ask the questions, wildcards can be used. This can be used without having any knowledge about the exact keywords and still high-quality structured information will be provided [7].

III. TEXT MINING TECHNIQUES:

Natural Language processing (NLP) [7] and Information Extraction (IE) [7][8] are the most popularly used techniques for text mining. There are other techniques that can be used for text mining such as knowledge-based, rule-based, statical and machine-learning based approaches. IE focuses mainly on extracting information from the actual text [10] while NLP focuses on text processing.
Advancement in technology has decreased the gap between human and computer understanding. The computer can also understand, analyze and generate text using Natural Language Processing. Some of the technologies are discussed below:

1) Natural language Processing (NLP)

Unstructured text information and automatic processing come under NLP. Its aim is to process the words found in the text. Two main fields are considered in it:

- Natural Language Generation (NLG): In order to make the generated text grammatically correct and fluent, NLG uses a linguistic representation of the text. Syntactic realizers are used by most of the systems to ensure that all the grammatical rules are efficiently followed. Machine translation system is one of the applications of NLG [10].
- Natural Language Understanding (NLU) [3]: The meaningful representations are found by NLU, by keeping a check on the discussions to the domain of computational language [11]. At least one of this constituent is present in it: tokenization, semantic analysis, morphological or lexical analysis and syntactic analysis.

A sentence is divided into a list of tokens by tokenization where these tokens represent a special symbol or word. Each word is tagged with its part of speech in the morphological and lexical analysis, it becomes complex when a word has more than one part of speech. Sentences can be broken into phrases, sub-phrases to the actual word by semantic analysis parse tree. There are two steps involved in semantic interpretations: a) Context Independent Interpretation: it looks for the meaning of words which helps to find the meaning of the sentence. b) Context Interpretation: it looks on the effects of the context of interpretation of the sentence. Context might include the situation of usage of sentences etc [7][11].

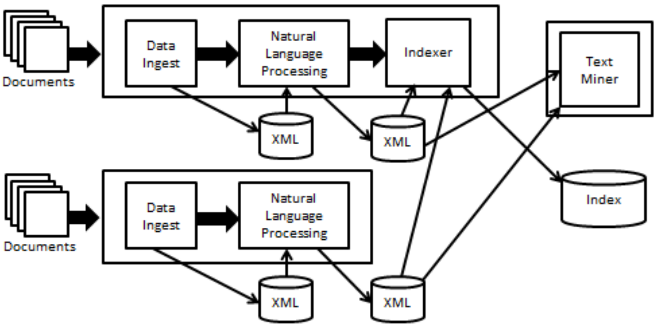


Figure 2: Natural Language Processing [7]

2) Information Extraction (IE)

Firstly, Extraction is done to analyze the structured text. The key phrases and relationships within the text are identified by designed software. Pattern matching is used for this purpose. The user can get the needed information by using the software that provides information about people, place and time. It becomes efficient when a huge amount of text is involved. The information that needs to be mined is in the form of relational database in traditional data mining whereas in other cases it is available in the free natural language [9]. In IE, the traditional data mining technique is applied to the corpus of textual document which is converted to a structured database.

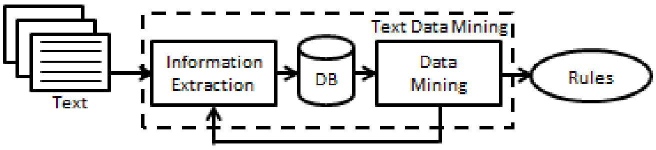
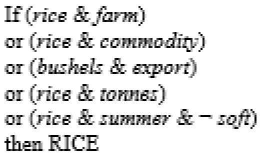


Figure 3: Overview of Information Extraction Text Mining [7][8]

Finally, to engage the most efficient induced rules, we filter the discovered rules on the basis of training and disjoint set. The changes that are made by IE on any validation templates or extracted training are discarded. Depending on whether the text appears on the document or not, the final decision of extracting or not extracting text is made. For example [7], RICE is to be extracted.



If (*rice & farm*)
or (*rice & commodity*)
or (*bushels & export*)
or (*rice & tonnes*)
or (*rice & summer & ¬ soft*)
then RICE

Figure 4: Rule for assigning a document to the category RICE

3) Topic Tracking

In this tracking tool, user interests are tracked and a document of interests of users is predicted based on their profile and documents viewed. Yahoo has provided a free tracking tool which notifies the user when some news related to the selected keyword appears. But there are drawbacks of this technology such as if an alert is a setup for data mining then it will show several stories and news related to mining the mineral rather than text mining.

It can be used by industries where the companies want to track their competitors in the market. It can also be used to track news of their own company and product [8]. The main process of topic tracking is keyword extraction. It includes a set of very important word that gives a complete description of the content of the article to the readers. It is very valuable to identify keywords from the huge amount of online news data because it can be used to make a small summary of a news article. Extraction of keywords manually is very difficult and time-consuming [9]. Some automated process is needed for fast extraction of keywords [7].

- Downloaded news documents are stored in the Document table.
- Dictionary table is used to store nouns extracted from documents.
- Term occur table consists of facts which word appear in documents.
- TF-IDF weights for each word is calculated by Term to occur fact table and the results are updates to TF-IDF weight table.
- Finally, TF-IDF weight table is used to rank the candidate keyword list for each domain with words.

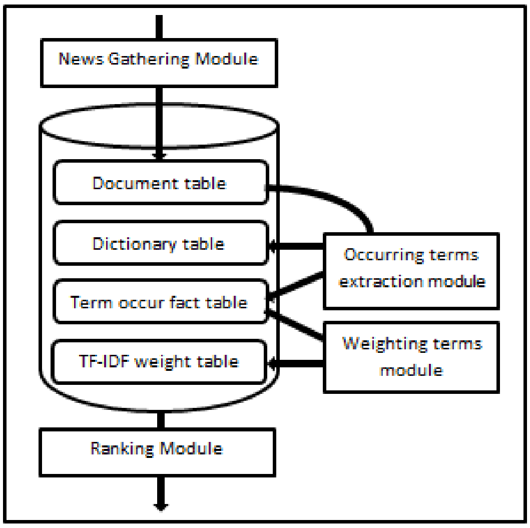


Figure 5: Keyword Extraction Module [7][9]

4) Summarization

Summarization of text is very important because it helps the reader to identify if the lengthy document is worth reading for supplementary information or not. Text summarization software is used in large text documents which summarize the document in the time user reads the first paragraph. The summarization process is based on the concept that it only decrease the length and details of the documents but the overall meaning and main points are retained. Although it is a difficult task to teach a computer to analyze semantics and interpret meaning [8].

As humans, we tend to read the complete document for full understanding and then write the summary with highlighted main points. Some alternative methods are used by computers because they lack language capabilities like humans. There are three steps in automated summarization process [9]:

- A structured representation of original text is obtained in pre-processing step.
- The text structure is transformed to the summary structure in processing step.
- The final summary is obtained from a summary structure in generation step.

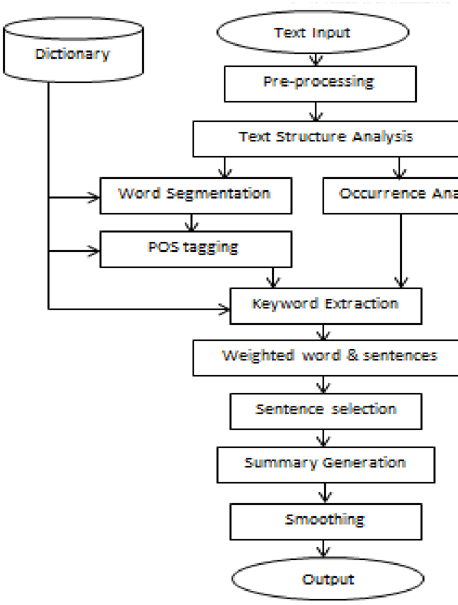


Figure 6: Text Summarization Process [7]

5) Categorization

In this documents is placed in a pre-defined set of topics in order to determine the main themes of documents. The computer treats the documents as a bag of words and actual information is not processed but only the words that appear are counter and main topics covered by the documents are identified. It relies on thesaurus and relationships can be identified by looking into narrow terms, broad terms, synonyms etc. In order to find the relevance of a document for a person finding information on a particular topic, topic tracking and categorization can be used together. From labeled documents classifiers can be learned by using supervised learning algorithm and classification is performed automatically on unlabeled documents.

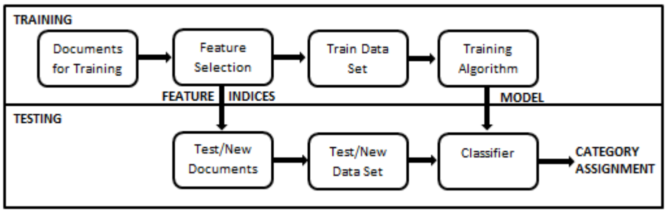


Figure 7: Categorization Process [7]

6) Clustering

In this similar documents are sorted into groups but it is different from categorization because the documents are clustered on the fly instead of predefined topics. One of the benefits of this is that documents can appear [4] in multiple subtopics without omitting the useful documents from the search results. A vector of topics is created by the clustering algorithm for each document and also measure the weight of how well documents fit into each cluster.

The high dimensionality of feature space is one of the problems with statistical text clustering.

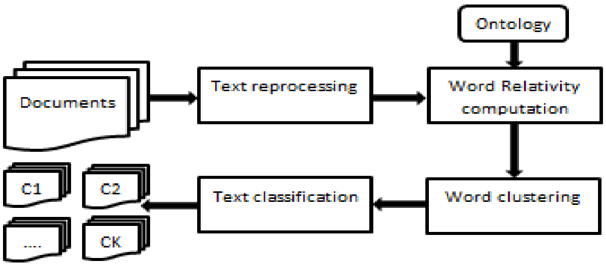
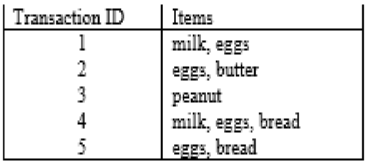


Figure 8: Clustering Process [7]

7) Association Rule Mining

It is used to find relations among a large set of variables in the data set. Industries have a huge advantage from this, it helps to discover the relationship among a large set of variables. Frequently occurring variable values are also checked by this technique. A relationship can contain 2 or more variables in ARM. ARM [4] is used to study relationships among topics in text mining. ARM finds the relationship among interesting correlation and association of large set of data items. ARM can be represented in form of mathematical representation as follows:



Transaction ID	Items
1	milk, eggs
2	eggs, butter
3	peanut
4	milk, eggs, bread
5	eggs, bread

- Support of A = {milk, eggs} = 2/5 = 0.4 = 40%
- Support of B = {bread} = 3/5 = 0.6 = 60%
- Support of A ⇒ B = 1/5 = 0.2 = 20%
- Confidence of A ⇒ B = $\frac{\text{Support of A} \Rightarrow \text{B}}{\text{Support of A}} = \frac{0.2}{0.4} = 0.5 = 50\%$
- Lift of A ⇒ B = $\frac{0.2}{(0.4)(0.6)} = 0.83$

IV. TOOLS OF TEXT MINING:

There are multiple tools present which can also be used to mine the text from the data inputted to them which also uses the techniques which we reviewed above. All these tools use a single pattern of steps to perform the data for example when a data event is to be analyzed, the first step is to send all the data into the mining tool. Once the data is gathered the linguistic processing is applied following with the Factor and cluster analysis of the data. The architecture is given as follows:

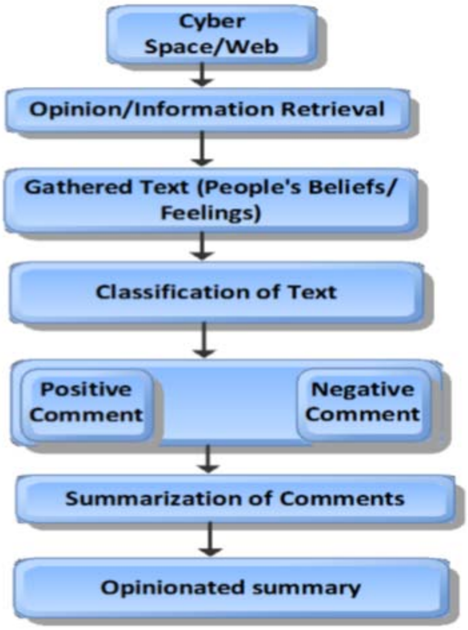

Cyber Space/Web
Opinion/Information Retrieval
Gathered Text (People's Beliefs/ Feelings)
Classification of Text
Positive Comment
Negative Comment
Summarization of Comments
Opinionated summary

Figure 9: Architecture of Opinion Mining [13]

Now we will study about some of the tools available for text analysis:

i. Megaputer Intelligence:

It is also called as the Text analyst which is based on natural language text analysis. Its mining framework applies loads of diagnostic capacities to explore content. To fabricate this tool 20 years of research is utilized. The fundamental preferred standpoint of this framework is it can distil the semantic system of a content totally based on its own experience without the help of a subjec tparticular lexicon by a human master [12].

ii. Intelligent Miner:

It was created by IBM and is a mixture of various analysis tools created by IBM only. The language detection tool, feature extraction tool, and all the other such tools are used to perform the task [12].

iii. Text Finder:

Created by Parcel Inc. is a fast and accurate text finder. It is one of the tools which can analyze large volumes of textual data. It is also useful for time constrained problems because it can analyze 0Mbyte text/second which is equivalent to 5000 pages.

iv. Callable Personal Librarian (CPL):

The people who are handy with C language can use this tools efficiently. The main advantage of the tool is that it can manage three different type of data i.e. structured, hypertext and multimedia. It helps in discovering models on the fly without external help [12].

v. Vantage Point:

It was used for competitive and technical intelligence. The system starts by conducting a search on the database using custom search engine created by the tools only. The preprocessing takes place which also includes the raw data. Lastly using NLP knowledge is extracted [13].

vi. Fulcrum Search Server:

It is basically a scrape which can scrape data from e-commerce website, customer care portals etc. The technique used for analysis is NLP, therefore, the output is reliable and accurate [13].

vii. Isaac and Amberfish:

Created by Etymon, Isaac is the free version and Amberfish is the commercial version of the tool. The search technique is open source which opens up for Boolean queries and field base searching too. A web scraping and searching are also provided in this tool [13].

viii. ISIS:

The UNESCO (United Nations Education, Scientific and Cultural Organization) is the creator of the tool. Like IBM this software is also an amalgamation of various software. The main advantage is that it has a strong retrieval engine and a lot more powerful formatting language is utilized. This tool is free if not used for commercial purpose [13].

ix. AE1:

AE1 is the searcher produced by Answer Logic which translates archives and inquiries by utilizing NLP methods. It first breaks down content by utilizing dialect processor and after that stores comes about as ideas in IdeaMap [13]. There is a report chief which deals with every one of the capacities in regards to archives. To get yield dialect processor coordinates the idea of question with put away ideas in IdeaMap and give best-coordinated ideas.

x. Wordsmith Tools:

Oxford university press produced the Wordsmith which analyzes the words from large chunks of text. It was based on three functions [13]:

a. Wordlist: it is the list of word cluster available for the used in alphabetical order.
b. Concord: it is the searcher helps in finding word or phrase from the text.
c. Keywords: it helps in identifying the keywords of the documents.

xi. Harvest:

It is planned by Internet Research Task Force Research Group. It is a blend of tools to chip away at the web and can accumulate, remove, arrange and repeat information. It is main stream for its quality to work in various configurations on various machines [13].

V. ANALYSIS:

In this section of the paper, we are going to analyze the technique which we studied in the paper above. Firstly we will start by differentiating between the classification, categorization, and clustering [14]. Since each of these techniques can be used in different fields based on need.

Technique	Features	Algorithm	Processes and models used
Classification	Efficient Accurate Simple Computation Suitable for continuous data Long Training is required	Support Vector Machine Decision Trees K-Nearest Neighbours Naïve Bayes Neural Networks Association rule-based Boosting	Data pre-processing Definition of training set and test sets Creation of the classification model using the selected classification algorithm Classification model validation Classification of new /unknown text documents.
Categorization	Adaptive Cost effective Moderately accurate Complex problem domain Other methods are required for knowledge extraction	Naive Bayes, SVM Neural networking Decision Tree K-Nearest Neighbour	Automatic: Typically exploiting machine learning techniques Vector space model based Prototype-based (Rocchio) Neural Networks (learn non-linear classifier) Support Vector Machines(SVM) Probabilistic or generative model based
Clustering	Proactive risk avoidance Adaptability Cost Effective Unsupervised learning Accurate Less test case needed Knowledge extraction is done by itself	Sequential algorithms Hierarchical algorithms Agglomerative algorithms Divisive algorithms Fuzzy clustering algorithms	Data pre-processing, remove stop words, stem, feature extraction, lexical analysis etc., Hierarchical clustering-compute similarities applying clustering algorithms. Model-Based clustering(Neural Network Approach) – clusters are represented by exemplars(e.g: SOM)
Summarization	Best for discontinue data Good for large amount of data Slow Moderately Accurate	Keyphrase Extraction TextRank LexRank PageRank KEA ,ROUGE GRASSHOPPER	Naïve Bayes Model Tools Used: Tropic Tracking Tool and Sentence Ext Tool

Table 1: Analysis of classification, categorization, and clustering based on models and algorithms.

In the above table we can see that the system which uses classification can use SVM, decision Tree, Boosting etc. to search the text from the data. The data here is pre-processed and various models based on the algorithms are created. The main advantage is that classification can help in finding knowledge from new/unknown data also.

Categorization, as shown, is a machine learning technique based on algorithms like Naive Bayes, SVM, Neural networking, Decision Tree, K-Nearest Neighbour. The system is more of a prototype based using Probabilistic or generative models. Lastly, clustering uses Sequential algorithms, Hierarchical algorithms, Agglomerative algorithms, Divisive algorithms, Fuzzy clustering algorithms and used hierarchical based or neural network based models.

VI. CONCLUSION:

In this paper, we have studied multiple methods which help in extracting knowledge from the text. The technique which came out to be best suited for the system is clustering, since the technique is adaptable, cost effective, has a good accuracy and is not complicated while applying the technique. Apart, from the standard parameters, the clustering also provides proactive risk avoidance, unsupervised learning leading to less number of test cases and finally the algorithm is dependent that is it doesn't require any external technique for knowledge extraction. Therefore we conclude that clustering is one of the best techniques for knowledge extraction with textual data.

VII.REFERENCES

[1] Ramzan Talib, Muhammad Kashif Hanif, Shaeela Ayesha, and Fakeeha Fatima, "Text Mining: Techniques, Applications, and Issues", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7 No. 11, 2016.

[2] R. Sagayam, "A survey of text mining: Retrieval, extraction and indexing techniques", International Journal of computational Engineering Research, vol. 2, no. 5, 2012.

[3] N. Padhy, D. Mishra, R. Panigrahi et al., "The survey of data mining applications and feature scope," arXiv preprint arXiv:1211.5723, 2012.

[4] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," Communications of the ACM, vol. 49, no. 9, pp. 76–82, 2006.

[5] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text mining: predictive methods for analyzing unstructured information. Springer Science and Business Media, 2010.

[6] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011," Expert Systems with Applications, vol. 39, no. 12, pp. 11303–11311, 2012

[7] Abhilasha Singh Rathor and Dr. Pankaj Garg, "Analysis on Text Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 2, February 2016.

[8] Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, August 2009.

[9] Naresh Kumar Nagwani, Dr. Shrish Verma "A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 17– No.2, March 2011.

[10] Minakshi R. Shinde1, Parmeet C. Gill, "Pattern Discovery Techniques for the Text Mining and its Applications", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Impact Factor (2012): 3.358 Volume 3 Issue 5, May 2014.

[11] 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.

[12] Jan van Gemert, "Text Mining Tools on the Internet: An Overview".

[13] Abhishek Kaushik and Sudhanshu Naithani, "A Comprehensive Study of Text Mining Approach", IJCSNS International Journal of Computer Science and Network Security, VOL.16 No.2, February 2016.

[14] R. Balamurugan, Dr. S. Pushpa, "A Review On Various Text Mining Techniques And Algorithms", 2nd International Conference on recent innovations in science, engineering, and management JNU convention center, New Delhi, 22nd November 2015.