**Database Emerging Technology Review**

**Text Mining – Techniques, Methods and Tools**

**Assignment 2**

**By:**

**Sanika Deshpande**

**Student ID : 805600277**

**CPSC 531-01 (13293)**

**Fall, 2016**

**Professor: Dr. Chun-l Phillip Chen**

**Department of Computer Science**

**California State University, Fullerton**

**Oct 27, 2016**

**Table of Contents**

Abstract ..................................................................................................................................................5

1. Introduction......................................................................................................................................6

1.1 What is Text Minning?.........................................................................................................6

1.2 Process of Text Mining.........................................................................................................6

2.0 Literature Review .............................................................................................................................7

2.1 Methods used in Text Mining .............................................................................................7

2.1.1 Term Based Method (TBM)………………………………………………………….7

2.1.2 Phrase Based Method (PBM) ………………………………………………………..7

2.1.3 Concept Based Method (CBM)………………………………………………………7

2.1.4 Pattern Taxonomy Method (PTM)…………………………………………………...8

2.2 Techniques used in Text Mining .........................................................................................8

2.2.1 Information Extraction (IE) ………………………………………………………….9

2.2.2 Topic Tracking ……………………………………………………………………...10

2.2.3 Summarization ………………………………………………………………….......12

2.2.4 Categorization ……………………………………………………………………....13

2.2.5 Clustering …………………………………………………………………………...15

2.2.6 Concept Linkage ……………………………………………………………………15

2.2.7 Information Visualization…………………………………………………………..16

2.2.8 Natural language Processing (NLP) ………………………………………………..17

2.3 Tools used in Text Mining……………………………………………………………….18

2.3.1 Sentiment Analysis ………………………………………………………………...18

2.3.2 Term Frequency – Inverse document Frequency (TF-IDF) ……………………….19

2.3.3 Named Entity Recognition (NER) …………………………………………………19

2.3.4 Event Extraction…………………………………………………………………….19

3.0 Discussion.......................................................................................................................................20

3.1 Advantages of Text Mining...............................................................................................20

3.1.1 Efficiency…………………………………………………………………………...20

3.1.2 Unlocking ‘hidden’ information and developing new knowledge …………….....20

3.1.3 Exploring new horizons…………………………………………………………...20

3.2 Challenges of Big Data Analytics in retail......................................................................20

3.2.1 Natural Language Issue …………………………………………………………...20

3.2.2 IE issues …………………………………………………………………………...21

3.2.3 Named entity recognition issue …………………………………………………...21

4. 0 Implications...................................................................................................................................22

4.1 Applications of Text Mining…………………………………………………………….22

5. 0 Current Trends and Future Trends.................................................................................................23

5.1 Current Trends……………………………………………………………………………23

5.2 Future Trends …………………………………………………………………………….23

6. 0 Conclusion………………………………………………………………………………………..24

References..............................................................................................................................................25

**List of Figures**

1. Figure 1: Process of Text mining …………………………………………………………..6

2. Figure 2: Text mining framework based on IE …………………………...........................9

3. Figure 3: Rule mining: pseudo code ……………………...………………………………10

4. Figure 4: Keyword Extraction System ……………………….........................................11

5. Figure 5: Keyword Extraction Module …………………………………………………...11

6. Figure 6: Summarization ………………………………………………………………….13

7. Figure 7: Text Categorization……………………………………………………………..14

8. Figure 8: Visualization…………………………………………………………………….16

9. Figure 9: Visualization Model called InfoVisModel …………………………………..17

**ABSTRACT**

Recently, increment in growth of data mining, digital data and knowledge discovery attracted attention for need of extracting useful information and knowledge from existing source. Challenge here is to extract user related information. Mainly text mining is used to extract information from unstructured to structured data. At first, this paper presents the framework of text mining that is process of text mining which includes five steps to derive the exact information needed by user. As text mining mainly deals with knowledge discovery, now it has very high commercial value in market. Text mining can also be used to extract hidden or new or previously unknown information. This paper mainly describes the methods, techniques and tools of text mining to solve challenged mentioned above. Discussion of these techniques and methods gives effectiveness over retrieval of information in text mining. This paper also highlighted challenges and current and upcoming trends of text mining.

**Keywords: Text Mining, Knowledge Discovery, Data mining**

1. **INTRODUCTION**
   1. **What is Text Mining?**

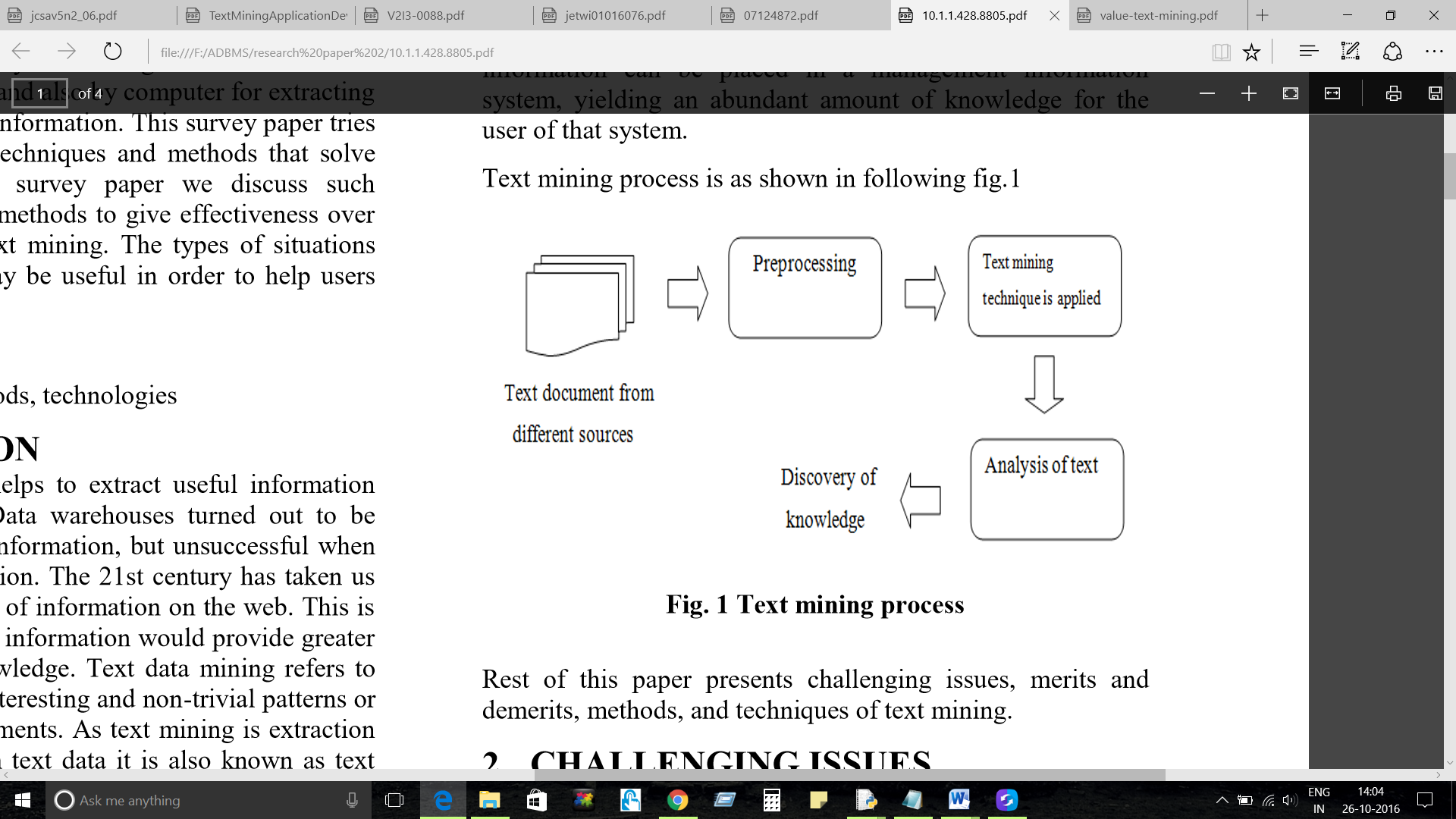
Equivalent to text analytics, Text mining makes unstructured data usable by computer. It filters large amount of research and extract information that user need.

**Definition of Text Mining:**

**“Text Mining is the process of deriving high-quality information from text”** (Luca Scagliarini, 2016)**.**

* 1. **Process of Text mining:**

Process of text mining begins with collection of documents from different sources. Then required document will get retrieved using text mining tool. It requires pre-processing after its retrieval. Pre-processing can be performed by checking its format and character set. Then next phase is of Text analysis in which semantic analysis is performed. Semantic analysis is performed on this document to extract high quality information from text. At this stage depending on organization’s goal, different combination of different text mining techniques is used. Sometimes these techniques works repeatedly until exact required information is extracted from text. (Monika D. Khatri, 2014) At last Management information system is used to place the resulted information which yields lot of knowledge.



**Figure 1 : Process of Text mining (Sonali Vijay Gaikwad, 2014)**

1. **LITERATURE REVIEW**
   1. **Methods used in Text Mining:**

Traditionally, many techniques are developed for relevant information retrieval according to user’s requirement that is to solve problem of text mining. Basically, there are four methods used for the same as follows:

* + 1. **Term Based Method(TBM) :**

‘Term’ is defined as ‘word having semantic meaning’. (Sonali Vijay Gaikwad, 2014) The document is analysed on the basis of term in this method. For term weighting, it has efficient computational performance and mature theories advantage. This method suffers from the problem of polysemy and synonymy. A word having multiple meanings called polysemy and a word having same meaning called as synonymy. It is difficult to retrieve semantic meaning of many discovered terms to answer what user wants.

* + 1. **Phrase Based Method(PBM):**

‘Phrase’ defined as information that is which carries more semantics. Phrases are more discriminative and less ambiguous than individual terms. In this method, document is analysed on basis of phrase. Reasons for daunting performance here include following:

1. Phrases contain inferior statistical properties to terms
2. Low frequency of occurrence
3. Redundant and noisy phrases are present in large amount.
   * 1. **Concept Based Method (CBM):**

Here terms are only analysed but on sentence and document level. The importance of word is captured by statistical analysis of term frequency. Frequency of two term can be same in same document but one term contributes more appropriate meaning than other term. This method is introduced to give more importance to the term that captures the semantics of text.

Three components are included in this model: first, Components which analyses sentences having semantics. Secondly, Components constructing a COG (conceptual ontological graph) used to derive the semantics. Third, to build feature vectors using standard vector space model, components that extract previous concepts are used. (Sonali Vijay Gaikwad, 2014) This model differentiates easily between terms which are not important and the terms deriving meaning of sentences. To optimize representation and to remove noise and ambiguity, the query concepts are used which applied by feature selection.

* + 1. **Pattern Taxonomy Method(PTM):**

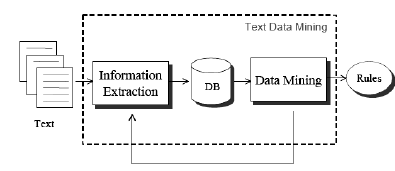
Now here, pattern takes into consideration while analysing the document. By using is-a relationship, patterns are used to structures into taxonomy. Pattern of information can be identified by using many techniques that uses data mining. However, low frequency problem can occur here that is less use of useful long pattern with high specificity. The problems called low-frequency and misinterpretation get reduced by using effective pattern discovery technique. Refinement of discovered pattern from text document can be done using this same technique. Hence, the final experimental results proved that this model performs better than all other models mentioned above.

* 1. **Techniques used in Text Mining**

For removal of gap between computer languages and human beings, technological advances are introduced. This technology is defined as natural language processing which mainly teach this language to computers to perceive, analyse and produce text. Technologies that use text mining and their job will be discussed in following sections. Also there will be illustration of situations where each of these technologies becomes useful for users to identify tools of their or their organization’s interest.

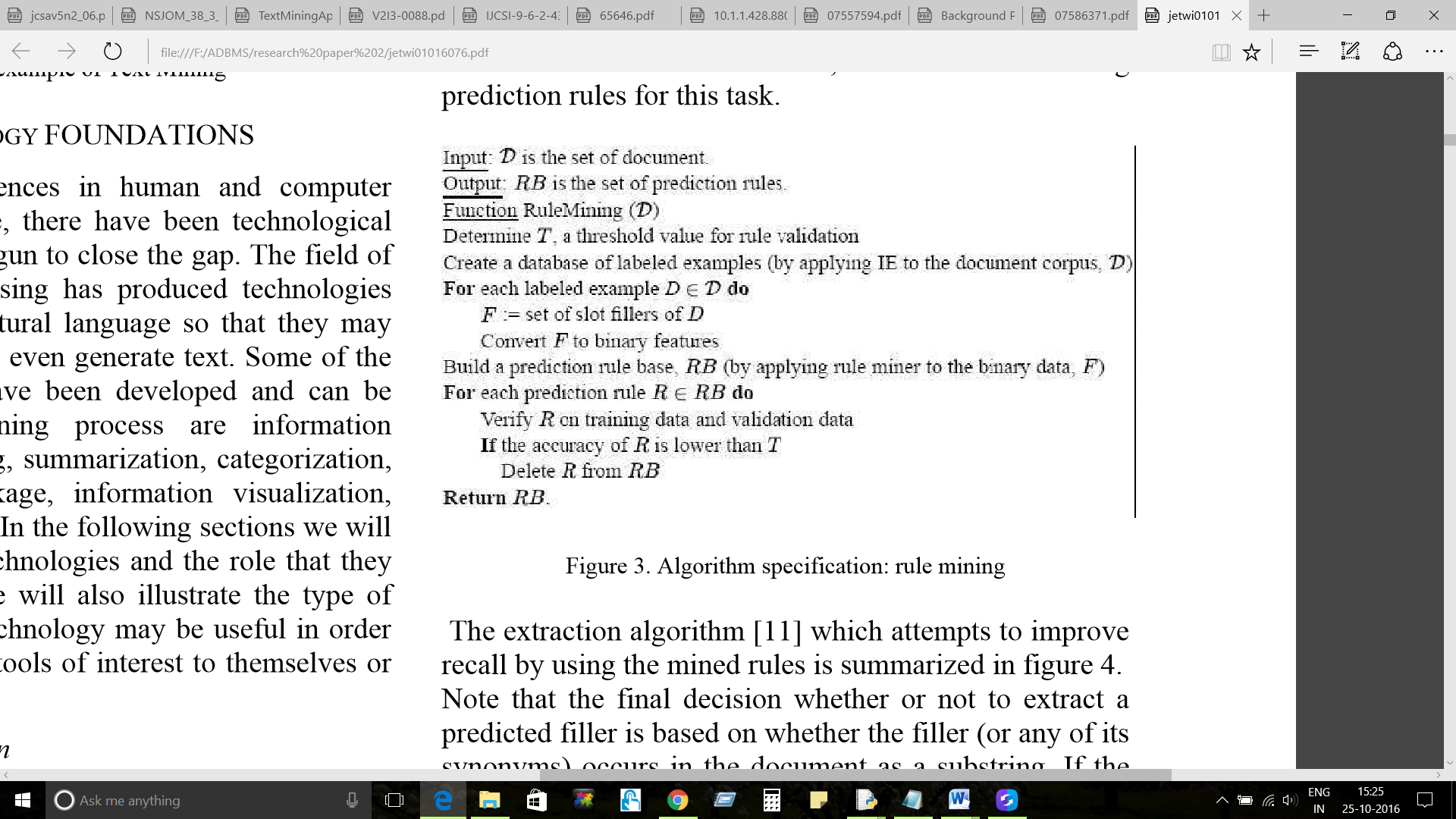
* + 1. **Information Extraction (IE)**

Information extraction (IE) is the start point of computers to analyse unstructured text. Pattern matching is the process in which IE software looks for predefined sequences named as key phrases or relationships within text. The process starts with semantically interpretation and parsing of phrases and sentences and then the resulted information entered into database. When dealing with large volume of text, it is very important to provide meaningful information to the user by inferring the relationships between recognized people, time and places. (Shaidah Jusoh, 2012) For mining purpose, IE transforms a mass of textual documents into structured database and then provides it to KDD module. This process is illustrated in following diagram.



**Figure 2: Text mining framework based on IE (Vishal Gupta, 2009)**

It is very necessary to mine the given rules for further extraction. The following figure of pseudo code is provided in which the final step of refining derived rules for retaining prompt rule. Templates will get discarded on which IE make any incorrect prediction. (Shaidah Jusoh, 2012) Instead of association rule and classification rule, here the prediction rules are focused that can be mined.

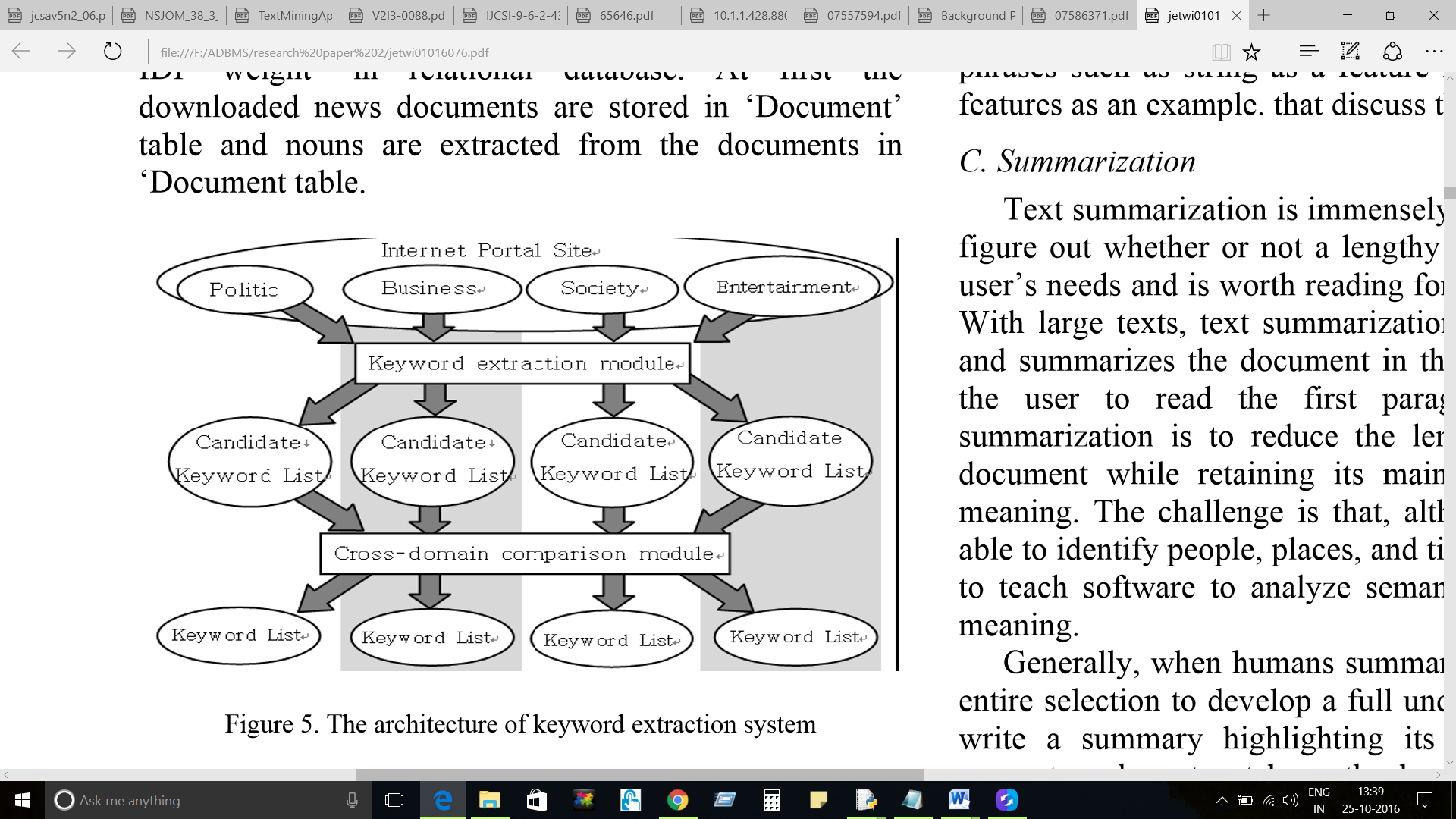


**Figure 3: Rule mining: pseudo code (Vishal Gupta, 2009)**

* + 1. **Topic Tracking**

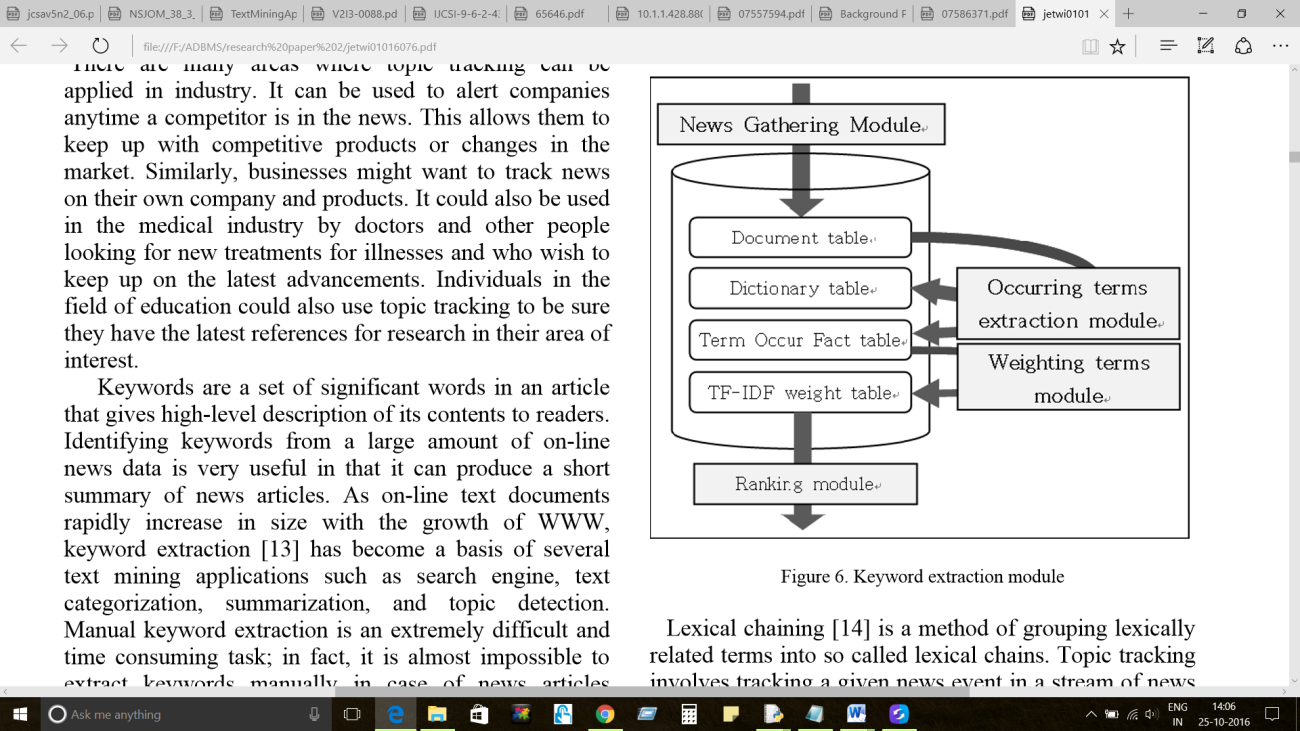
Here, prediction of the documents which user may need in future is done by this system. For this prediction, system store profiles of users and analyses document which user mostly sights. For example, topic tracking tool of yahoo is used by users to choose keywords and then that tool notify same user when news on related topics becomes available. (Vishal Gupta, 2009) Software can automatically infer the user’s interest by allowing them to choose the category of interests or by reviewing their history. Companies are utilizing topic tracking in many areas. When the competitor of any company is in news then this tool is used to alert that company about it. This becomes helpful for company to be updated about competitive products or changes in market. Sometimes not only about competitors, but company may want to keep track of their own news and make use of this news for their further advancements.

Collections of important words that exhale high level illustration of its information to its readers are called keywords. Now these keywords are useful in case of making summary of large amount of news articles. By identifying and collecting these keywords together, it has become very easy to create a short summary out of it. Keyword extraction is the main task which is done by almost every technique in text mining as there is a rapid increment in size of on-line text documents. This Keyword extraction is very time consuming and difficult if done manually. For rapid keyword extraction, it is necessary to establish an automated process. The following figure shows the steps of keyword extraction process used with HTML news pages.



**Figure 4: Keyword Extraction System (Vishal Gupta, 2009)**

Internet Portal Site has HTML pages representing news. Keyword extraction module extricates the possible keywords from that site. Then extraction of keywords is performed by cross domain compare son module. For this process, first we need to understand how keyword extraction module works. Following figure shows the steps and components involved in keyword extraction module.



**Figure 5: Keyword Extraction Module (Vishal Gupta, 2009)**

In this module, relational database contains tables like ‘document’, ‘dictionary’ , ‘TF-IDF’ and ‘term occur fact’. (Vishal Gupta, 2009) First, Document table contains downloaded news documents and nouns are getting extracted from this table. Then the ‘term occur fact’ table is getting updated by the facts. By using this table, calculation of TF-IDF weight for each word is accomplished and result of that are being updated to TF-IDF weight table. Finally, candidate keyword list is created which contains top ranking words for different domains by using ‘TF-IDF weight’ table.

* + 1. **Summarization**

In this technique, software retain most important points and general meaning by minimizing the document’s length with its details as it requires to figure out about meeting user’s needs by document which greater in length. Summary replaces the set of documents. User’s time get reduced when user read the summary of the set of documents instead of reading everything. Even though, computers have ability to recognize places, time and people, it is very arduous for software to read and analyse the semantics and interpret their meaning.

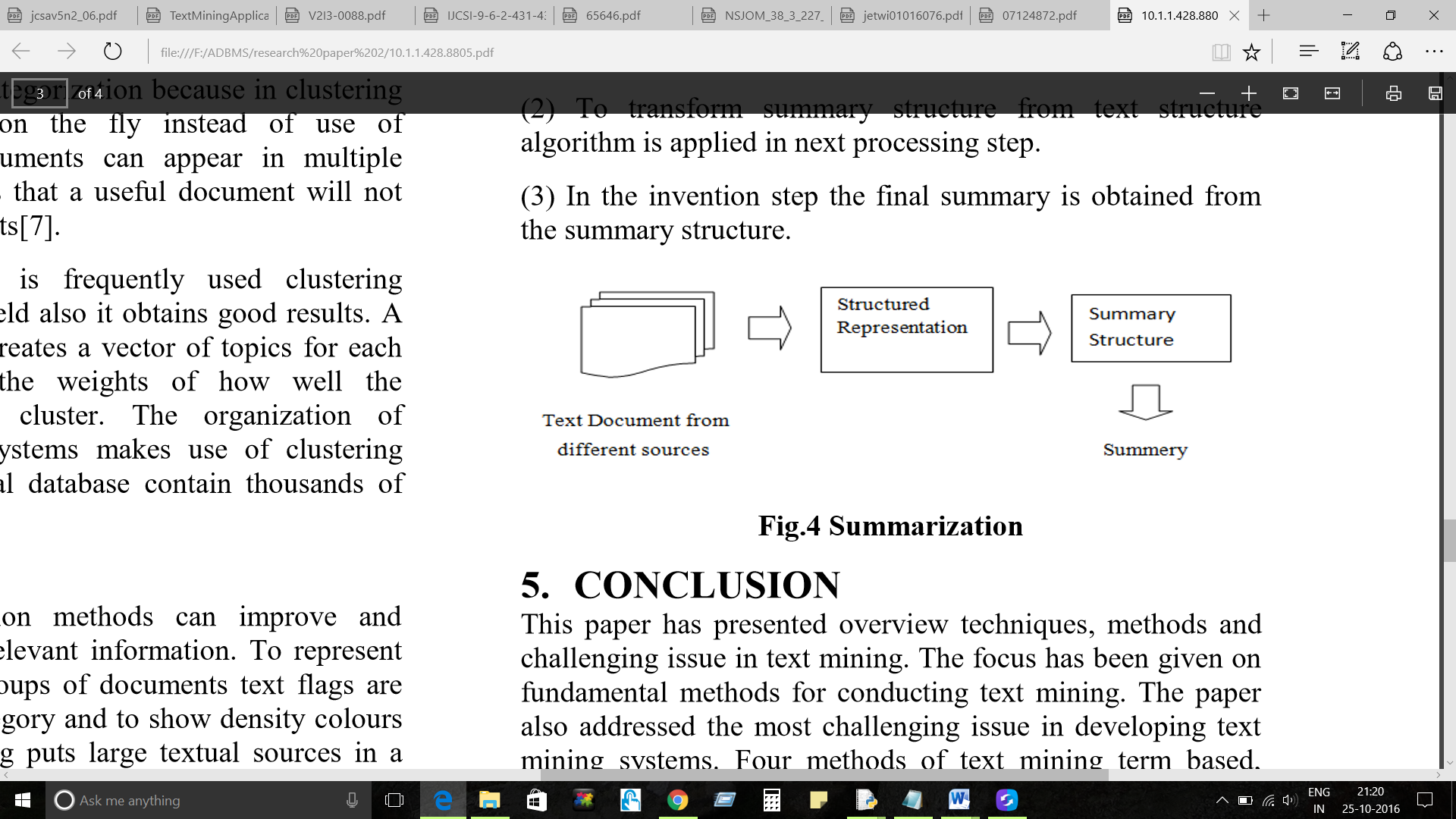
For humans, it is very easy to read entire text paragraph and then try to develop full understandable summary by highlighting it main points. (Sonali Vijay Gaikwad, 2014) So, for fulfilling this requirement by computer, the strategy used is sentence extraction from document which statistically weights the sentences and extracts the important sentences from it. The next strategy is called position information where for example, if the document contains words as “in conclusion”, then ultimately, after these words main points of documents will lie.

AutoSummarize function used in Microsoft word is simple example of text summarization. (Vishal Gupta, 2009) This tool allow user to choose the text which they want to summarize for their better understanding and limiting time.

Following are the steps of summarization process:

1. Pre-processing obtain a structured representation of original text
2. Application of transformation of text structure algorithm in summary structure.
3. Obtaining final summary from summary structure

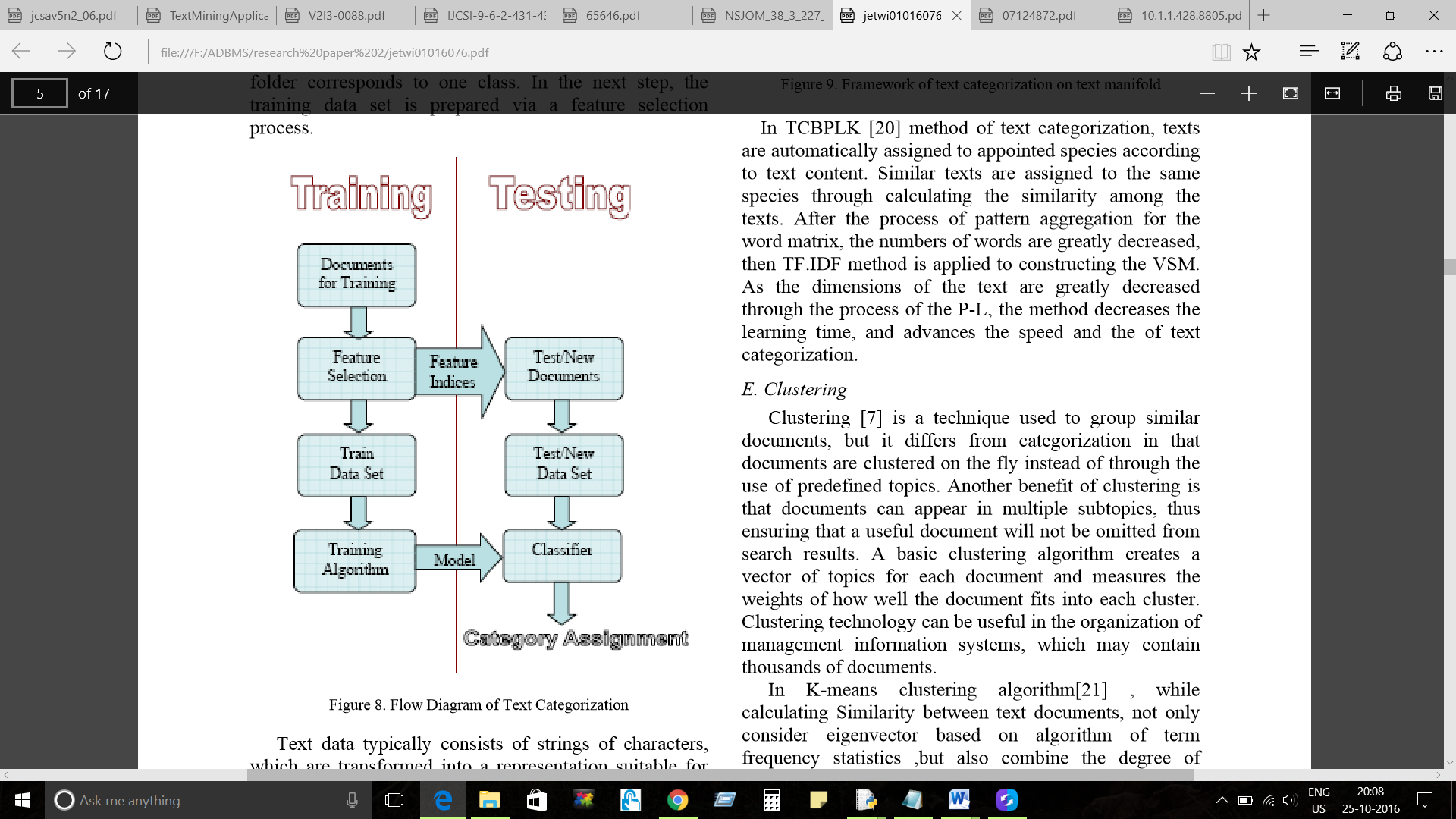
This last step of obtaining final summary is called as invention step.



**Figure 6: Summarization (Sonali Vijay Gaikwad, 2014)**

* + 1. **Categorization**

Categorization is defined as to ascertain the document’s main concept. It can be done by searching the predefined set of topics and placing the document into it. Instead if extraction categorization counts the words from document and identify the main topic of document. Relationships can be identifies from synonyms, related terms, broad terms and narrower terms. Thesaurus deals with predefined topics. So, Categorization relies on relationship and thesaurus. (Vishal Gupta, 2009) Ranking on the basis of the documents which have most content on specific topic is done by categorization. To get most relevant document at first, the ranking of document returned from topic tracking on the basis of weight of contents is done by categorization technique. This technique is used in many business applications in customer support area. In Supervised learning, algorithms are mainly used to learn the classifiers from labelled documents. These classifiers then used to implement classification operation on unlabelled documents.



**Figure 7: Text categorization (Vishal Gupta, 2009)**

In above figure, set of labelled documents (D) belongs to set of classes (C). Now here by using supervised learning algorithm will assign these trained classifier categories to new documents. (Sonali Vijay Gaikwad, 2014) Feature selection process is used for training data set. In this phase, different set of documents arranged to separate folders where each folder is a class. Process of text categorization includes indexing, pre-processing, classification and dimensionally reduction.

* + 1. **Clustering**

The process of finding the groups of documents with similar contents is called as clustering. The result of clustering is a partition (p) and numbers of documents (d) are present in each cluster. (Sonali Vijay Gaikwad, 2014) If one cluster contains most similar documents and most dissimilar documents are present in two separate clusters then the quality of clustering is consider better. K-means is clustering algorithm mainly used in text mining field to obtain good results. Clustering algorithm measure the weights of how well document fit into cluster.

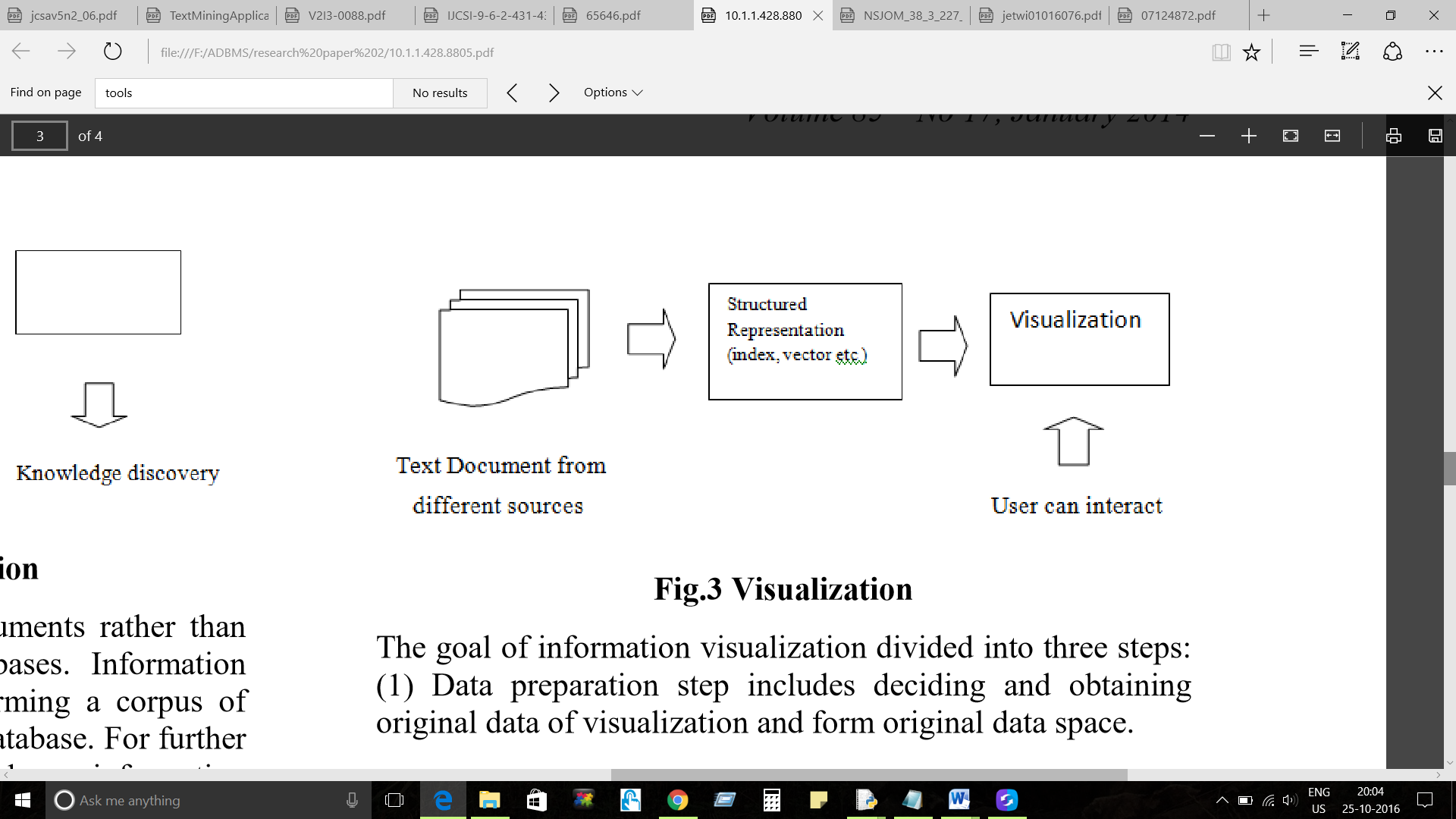
Steps of clustering text documents are:

1. **Stop word removal**: Stop words like pronouns, prepositions are those words which do not carry any information. So removal of these words will improve results of clustering.
2. **Stemming (Suffix Removal)** : Grouping of words having same conceptual meaning is very important such as work, worker , worked and working should be grouped by removing their suffixes.
3. **Filtering**: Reducing the dimensions of documents will reduce the problem of high dimensionality of feature space.
   * 1. **Concept Linkage**

Aim of this method is to connect related documents by helping users to find information which they wish to and identifying commonly shared concepts of documents. It focuses more on browsing for information than searching for it. (Vishal Gupta, 2009) For example, biomedical field – Identification of link between diseases and treatments can be done by this tool which humans cannot.

* + 1. **Information Visualization**

Discovery of relevant information is simplified by visualization. Document category and density colours can be shown by text flags which represents individual document or group of document. Large textual sources are put into a visual hierarchy by using visual text mining. Zooming and scaling are the main operations used by the users to interact with the document. (Vishal Gupta, 2009) For example, to identify terrorist network or crime information, government uses information visualization. Following diagram shows the process of visualization.



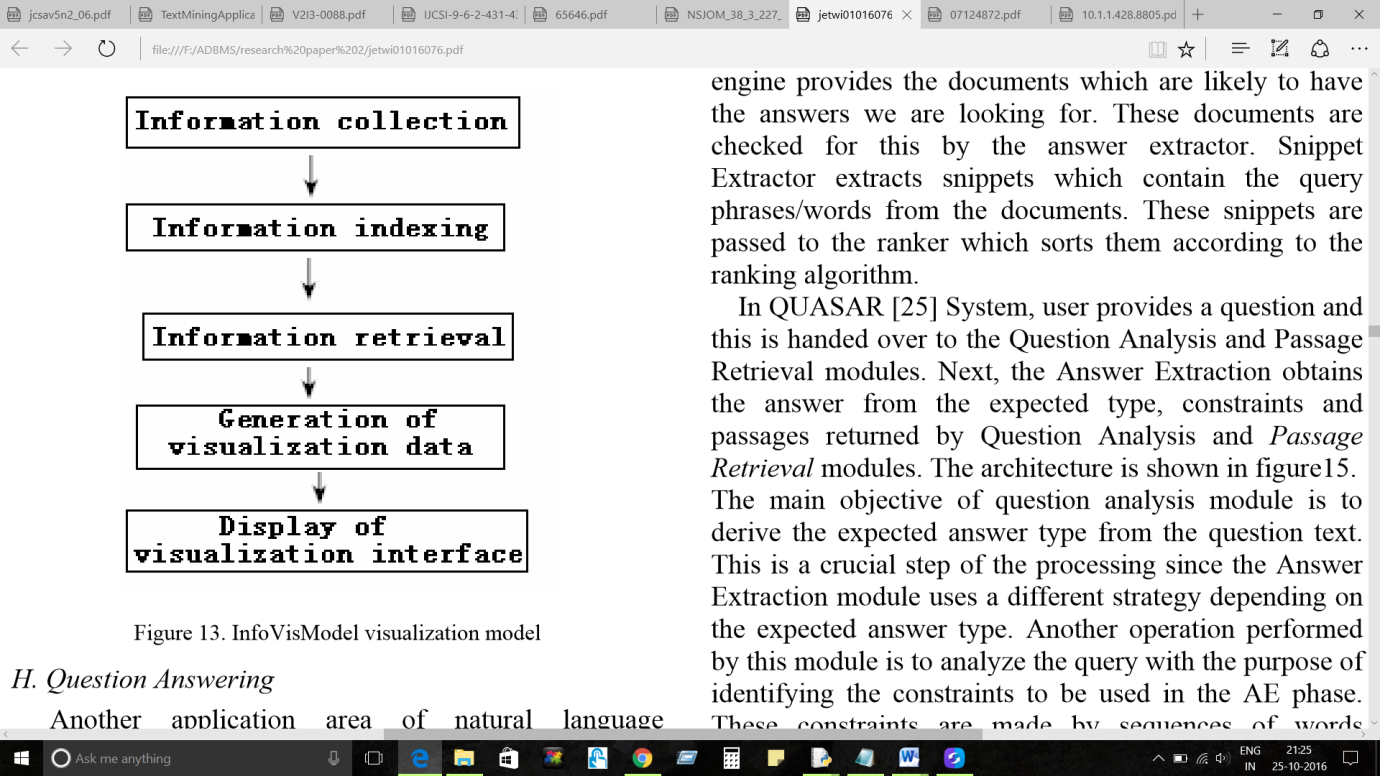
**Figure 8 : Visualization (Sonali Vijay Gaikwad, 2014)**

**Visualization consists of three steps:**

1. Obtaining original data of visualization and forming visualization data space is included in this first step defined as data preparation step.
2. In Data analysis and extraction step, tool form an visualization space by analysing and extracting visualization data from original data.
3. By using mapping algorithm, the visualization space is mapped to visualization target in this last Visualization mapping step.

**Construction of** **InforVisModel categorized into 5 steps:**

1. **Gathering Information**: Information should be accumulated from different sources.
2. **Creation of index from Information**: To form original databases, indexing of collected information resources has been done.
3. **Retrieval of Information**: Querying the information lists in conformity as and when required.
4. **Producing Visual data** : Retrieved data gets transformed into a form that can be easily visualized.
5. **Displaying visual interface**: In this step, visualized data mapped into visualized target and displayed it.



**Figure 9: Visualization Model called InfoVisModel (Vishal Gupta, 2009)**

* + 1. **Natural Language Processing (NLP) :**

Application of Natural Language Generation (NLG) called machine translation system is used to analyse text from source language and generate corresponding text in target language. By restricting the discussion to domain of linguistic, the Natural Language Understanding (NLU) is used to determine the meaning representation. Components included in NLU are tokenization, syntactic analysis, semantic analysis, lexical analysis and morphological. (Shaidah Jusoh, 2012) Dividing sentences into tokens where tokens are nothing but the word or special symbol. Tagging a word with its part of speech comes under morphological or lexical analysis. It has become complex in process of tagging words with various parts of speech. Semantic analysis is related to unambiguous sematic representation of syntactic structure of sentence. There are two steps in semantic interpretation:

1. **Context independent interpretation**: It determines the meaning of word and when these word will combine then what will be the meaning of sentence.
2. **Context interpretation**: It determines interpretation of sentences affected by context.
   1. **Tools used in Text Mining**
      1. **Sentiment Analysis:**

Customer reviews about the products and service or performance in market of company helps in responding faster to overcome the problems. There are total three analysis of sentiment analysis. First, Polarity analysis in which tone of communication that is positive or negative is identified. Second, Categorization where using fine-grained tools, it understands confusion of someone. It also put scale on emotion from 0-10 and from ‘sad’ to ‘happy’.

ANEW i.e. Affective Norms for English Word has been proven a significant tool for rating emotions. (Merrett, 2015). WordNet is other tool that is used to check relation between words that whether they are similar to each other or not like synonym or antonym and let users build classification schemes. Other free tools like Natural Language Toolkit and TextBlog is used for sentiment analysis. Also commercial tools like RapidMiner or ViralHeat are also used (Merrett, 2015).

* + 1. **Term Frequency – Inverse Document Frequency (TF-IDF) :**

It checks the number of occurrences of words in document and its importance relating to whole document. For example, words like prepositions can appear frequently in document but those are meaningless whereas words that appear frequently and its meaning is related to main topic of document then keeping track of those useful words is what TF-IDF captures. (Merrett, 2015) It is mainly used to build predictive model or classifiers. In case of classification, machine learning algorithms use the weights of the words that appear frequently.

* + 1. **Named Entity Recognition (NER)**

NER recognizes nouns and extracts person, dates, locations, organization or like from text. It performs this operation by looking at surrounding words. It also recognizes the pattern and lists the properties like uppercase, lowercase letter and punctuation that appear after word.

* + 1. **Event Extraction**

Event extraction is harder step than NER. It not only focuses on nouns but also find the relationship between them and find out inferences from incidents in text.

1. **DISCUSSIONS**
   1. **Advantages of Text Mining:**
      1. **Efficiency:**

A key benefit is to improve efficiency of analysis of knowledge. Time reduction because of automatic extraction method ensures coverage of domain knowledge. There is also a cutting down of time required to identify relevant material using text mining process.

* + 1. **Unlocking ‘hidden’ information and developing new knowledge :**

Unlocked hidden information provides some new knowledge and understanding. The large amount of academic publications and literatures if relates each other then there could be some underlying connections between subtopics mentioned in those documents which could only be analysed using automated analysis. For example, links found between diseases and drugs by unlocking hidden relation or information between them.

* + 1. **Exploring new horizons :**

Text mining not only concern with how research is done but it also concern with the fact that what is researched because of which new questions and horizons are explored. Emerging of new areas of digital humanities can be a good example to explain this.

* 1. **Challenges of Text Mining**
     1. **Natural Language Issue:**

It deals with ambiguity problem. Understanding in two or more ways means ambiguity. Ambiguity cannot be fully removed from natural language because it gives flexibility and usability. Also one word can have multiple meanings. After a lot of research, the result shows that ambiguity cannot be removed and work on this is still going on.

* + 1. **IE issues:**

More limited task has been performed by IE than full text understanding. Relations between the words presented can be completely delimited and text can be represented.

* + 1. **Named entity recognition (NER) issue :**

It identifies entity in free text. NER forms starting point. It means after recognition of correct entities search for its pattern and relation begins. Here also ambiguity problem is present. Ambiguity of entities has to be resolved. However, sense disambiguation concludes that this annotation ambiguous task is an expensive task.

1. **IMPLICATIONS**
   1. **Applications of text mining:**

Technology can now be applied broadly in various different fields. Lot of categories are there in which applications can be easily created. These categories are:

* Record Management
* Business Intelligence
* Listening Platforms
* Automated ad placement
* Publishing
* Responding to customer reviews in retail industry
* Information access

**Security Applications:**

It involves study of text encryption or text decryption. For national security purpose, online plain text has been analysed and monitored using many text mining software packages.

**Biomedical Application:**

PubGene is biomedical application that combines network visualization and biomedical text mining as internet service. (Kaushik, 2013) Also Go PubMed is a knowledge-based search engine for biomedical text.

**Software Applications:**

IBM and Microsoft use text mining to improve the results in the area of search and indexing. Specifically for monitoring and tracking terrorist activities the text mining software within public sector is created.

1. **CURRENT TRENDS AND FUTURE TRENDS**
   1. **Current Trends**
2. **Fight against terrorism:**

Increasing terrorism leads to major attacks in almost all countries. It helped to implement a law of fight against terrorism. Many problems faced by these countries while launching various programs are:

* These programs contained text, videos etc. are stored in database.
* Increment in size of data leads to increment in execution time

1. **Curing diseases with the help of Bio-informatics**

Humans should stay healthy to live long lasting life. To cure the diseases mainly text mining techniques are used. It predicts history of patient about diseases and can tell the current data for patient.

1. **Semantic web:**

Internet has become main part of human life like health only. Now days, human use internet like they can’t live without using it. Web contains lot of unstructured data that will get organized by semantic web.

1. **Business Trends:**

Two text mining techniques prediction and classification are used in business. Nowadays, larger needs are faster and accurate. These techniques are used to improve business and productivity of it.

* 1. **Future Trends**

1. **Distributed text mining:**

Data located only in one place in Data Warehouse or database can use by text mining technique. Future Scope in this is to use this mining technique on the data placed at various places. These distributed data at a local level will get combined to form global level data. However, combining data with different characteristics will create ambiguous problem.

1. **Phenomenal text mining:**

It mainly concern with the relation between data and phenomena. For example, by examining the data for phenomena or by implementing a program, information of customer that is customer’s details can reveal many things like purchase habits or age etc.

1. **CONCLUSION**

In this paper, I conclude that a process of extracting useful text from unstructured information by making it structured data for further processing on it is called text mining. Text mining has high potential value. Knowledge discovery Text (KDT) is mainly used to discover the knowledge from many sources. Semantic relation is placed between the words using NLP techniques. This paper also explained the tools which generally used in real world for text mining in different areas. Real time applications are mentioned in this paper where text mining is mainly used in the industry. In this paper the focus is mainly on techniques and fundamental methods for handling text mining. Four methods and eight techniques are explained in this paper which shows different ways of extracting and measuring the data in text mining for getting useful result. To provide improved understanding of text mining current trends and future trends are also explained which indirectly shows current and future scope of this paper respectively.

# References

Kaushik, M. L. (2013). Text Mining - Scope and Applications. *Journal of Computer Science and Applications*.

Luca Scagliarini, M. V. (2016). *The definition of text mining, it only seems simple but the real activity is not*. Retrieved from Expert system symantic intelligence: http://www.expertsystem.com/definition-of-text-mining/

Merrett, R. (2015). *5 tools and techniques for text analytics.* Retrieved from CIO from IDG: http://www.cio.com.au/article/575209/5-tools-techniques-text-analytics/?

Monika D. Khatri, S. D. (2014). International Journal of Advance Research in Computer Science and Management Studies . *History and Current and Future trends of Data mining Techniques* .

Shaidah Jusoh, H. M. (2012). Techn TTeecchhnn Techniq iiqq iques, Applications and Challenging Issue in Text Mining . *IJCSI International Journal of Computer Science Issues*.

Sonali Vijay Gaikwad, A. C. (2014). Text Mining Methods and Techniques . *International Journal of Computer Applications* .

U. Ackermann, B. A. (2014). SpeeData: Multilingual Spoken Data Entry. *IEEE.*

Vishal Gupta, G. S. (2009). JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE. *A Survey of Text Mining Techniques and Applications* .