Volume 6, Issue 2, February 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Analysis on Text Mining Techniques

Abhilasha Singh Rathor
Uttarakhand Technical University
India

Dr. Pankaj Garg
Dev Bhoomi Group of Institutions
India

Abstract— With the advancement in the field of computer science data has also become huge and hard to manage. Data Mining provides efficient ways of exploring and analyzing huge data to extract meaningful patterns and rules. In this paper we aim at discussing one of the important fields of data mining i.e. Text Mining and its techniques. Text is relatively cheap but extracting information about what text belongs to is expensive. The choice of finding particular technique to apply in a particular situation depends on nature of text mining task, the nature of available text, and the skills and preferences of data miner.

Keywords— Data Mining, Text Mining, Topic Tracking, Opinion Mining, Information Extraction

I. INTRODUCTION

Data Mining provides efficient ways of exploring and analyzing huge data to extract meaningful patterns and rules. Data mining algorithms were invented firstly for commercial applications such as for improvement of sales, marketing and customer support, but they are equally applicable in other fields also. Building models is the main concern of Data mining to connect inputs to particular output. Data mining is broadly of two types- directed and undirected [1].

1) Directed Data Mining aims at categorizing or explaining some particular field. It is top down approach and user already knows what he is looking for.

2) Undirected Data Mining is bottom up approach where data speaks itself. Patterns are found in data and it is left to the user to decide whether these patterns are important or not.

A. Issues and Challenges

Data mining is not easy to perform since for different type of data different complex algorithms are used and data is also not available at a single place. It needs to be collected from many diverse sources. So data mining is one of the challenging fields with following issues [2]:

a) Data available for data mining is very huge, large set of variables and cases may appear. Size of data may vary from multi gigabyte to terabyte. So there is need of deciding efficient algorithm for processing the data.

b) The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from heterogeneous databases adds challenges to data mining.

c) The database contains complex data objects, spatial data, multimedia data objects, temporal data etc. It is hard for one system to handle relational and complex data.

II. DATA SOURCE

There are many data sources for text mining. We can perform text mining on papers, news articles, patents or databases. Text mining can also be performed on web sources like social media data, Wikipedia, data repositories etc. Figure 1 shows different text mining sources.

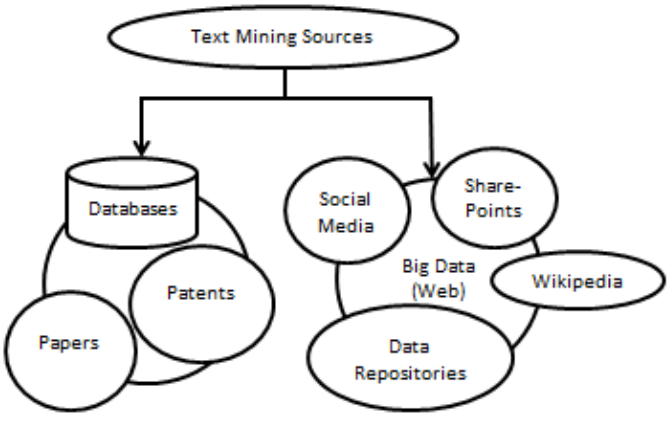

Figure1: Text Mining Sources [1]

III. TEXT MINING

Text is most common way of exchanging information. Although extracting information from a text document is not an easy task, it requires intelligent tools which can extract information very quickly and efficiently at low cost. So text mining produces intelligent tools to extract useful information from unstructured text document. The tool is a text mining system which has the capability to analyze large quantities of natural language text and detects lexical and linguistic usage patterns in an attempt to extract meaningful and useful information [3].

Data Mining is looking for patterns in data while text mining is looking for patterns in text. The real difference between the two lies in its superficial similarity. Data mining can be characterized as the extraction of previously unknown, implicit, and potentially useful information from data. The information in the input data is unknown, hidden and can be hardly extracted without using explicit data mining techniques. While in text mining the information to be extracted is clearly known and explicitly stated in text. The information is not hidden as author's express them clearly from human point of view [4]. The problem is that information does not lie in a structured manner so text mining techniques are required to extract text in a form which is easily understandable by user and there is no need for human intervention. From computers point of view both data mining and text mining problems are quite similar.

A. Working of Text Mining

a) Predefined keywords are searched from the documents using traditional search techniques, after that more accurate information is extracted such as relationships, concepts, sentences, phrases and even numerical information such as age, year etc.

b) To extract meaning of text, identify, synthesize, extract and analyze relevant fact and relationships text mining software tools are used. These tools use computational algorithms based on Natural language Processing (NLP), to enable computer to read and analyze textual information.

c) Text is mined in comprehensive, reproducible and systematic way so that needed information can be captured repeatedly.

d) Pre-written influential NLP based queries can run in real time across millions of documents.

e) Wildcards can be used to ask questions which we are looking for, without having to know the exact keywords and still get the high quality structured information.

B. Techniques

There are many techniques used for text mining, but most popular are Natural Language Processing (NLP) and Information Extraction (IE). Research is going on to analyze other techniques for text mining such as knowledge based, statical, rule-based and machine-learning-based approaches. NLP focuses on text processing while IE focuses on extracting information from actual text [8]. After this the extracted information can be stored in databases, further data mined and summarized. NLP techniques enable text mining tools to get closer to the semantics of a text source [3]. This is important, especially when the text mining tool is expected to discover knowledge from texts.

There is large difference between human and computer understanding, but advancement in technology has decreased the gap. Natural Language processing field has produced the technologies that teach computer natural languages so that computers can also analyze, understand and generate text. Some of these technologies are discussed here so that readers can identify tools of their interest and need.

1)Natural language Processing (NLP)

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

- Natural Language Generation (NLG): NLG uses linguistic representation of text, so that generated text is grammatically correct and fluent [4]. Most of these systems can include syntactic realizers ensuring that grammatical rules are efficiently followed or not. One example of NLG application is machine translation system [8].
- Natural Language Understanding (NLU): NLU is a structure that finds the meaningful representation, by checking the discussion to the domain of computational language [9]. It contains at least one of these constituents: morphological or lexical analysis, tokenization, semantic analysis and syntactic analysis.

Tokenization divides a sentence into list of tokens, which represents a special symbol or word. In morphological or lexical analysis each word is tagged with its part of speech, it becomes complex when a word contains more than one part of speech [5]. Syntactic analysis assigns parse tree to given natural language sentence, determining broking of sentence into phrases, sub phrases to actual word. In Semantic Analysis syntactic structure of sentence is translated into semantic representation, which allows system to perform appropriate task in its application domain [4][9]. Semantic interpretation contains two steps: a) *Context Independent Interpretation:* that concerns with the meaning of words and combining these meanings into sentences to find meaning of sentences. b) *Context Interpretation:* it concerns with effect of context on interpretation of sentences [9]. Context includes situation of usage of sentence, preceding sentences etc.

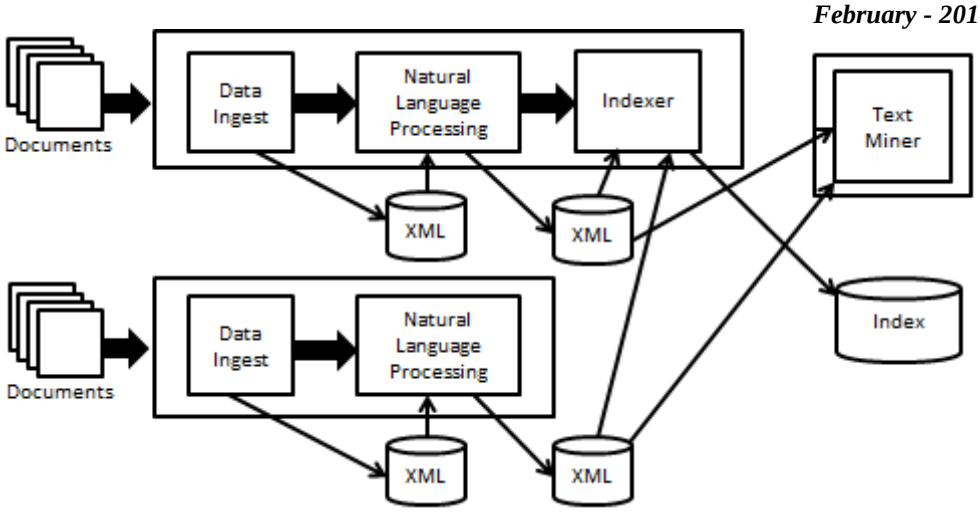

Figure 2: Natural Language Processing [4]

2)Information Extraction(IE)

To analyze structured text first of all information extraction is done. Software designed for this purpose identifies key phrases and relationships within the text [6]. This is done by using pattern matching. Software find out people, places and time to provide user with needed information. It is very efficient when we are dealing with huge amount of text. In traditional data mining the information to be mined is in the form of relational database, but in most of the applications the information is only available in the form of free natural language [7]. In IE the corpus of textual document is firstly converted to structured database and then traditional data mining techniques can be applied.

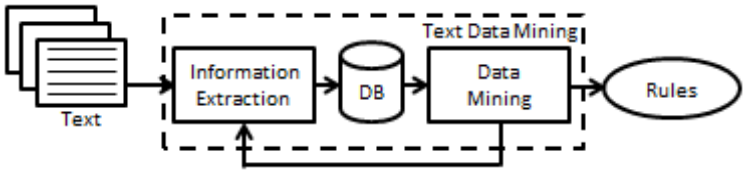

Figure 3: Overview of Information Extraction text mining [6]

Final we filter the discovered rules on the basis of both training data and disjoint set so that most efficient induced rules are engaged. The changes made by IE on any extracted training or validation templates are discarded. The final decision of extracting or not to extracting text is based on whether the text occurs in the document or not [6]. For example 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 document of interests of users are predicted based on user profile and documents viewed by user. Free topic tracking tool is offered by yahoo (www.alerts.yahoo.com) which notifies users when news related to chosen keyword becomes available. But this technology has its own limitations, taking example if we setup alert for "data mining" then we will receive several news and stories related to mining the minerals rather than text mining.

Topic tracking can be significantly applied in industries where companies can generate alert to find competitors in news. They can also track news of their own product and company. Keyword extraction is main process in topic tracking; it is set of important words in an article which gives high level description of its content to readers [6]. Identifying keywords from huge amount of online news data is very valuable, since it can create short summary of news article [7]. Manual extraction of keywords is very difficult and time consuming. For fast extraction of keywords automated process is needed. Keyword extraction system is shown in figure 4 [6].

- 'Document Table' stores downloaded news documents.
- 'Dictionary table' stores nouns extracted from documents.
- 'Term occur table' contains the facts which words appear in documents.
- 'Term occur fact table' calculates TF-IDF weights for each word, and the results 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 [7].

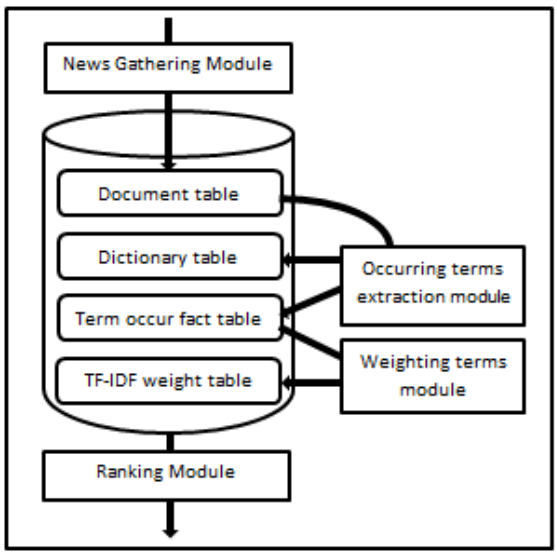

Figure 5: Keyword Extraction Module [7]

4) Summarization

Summarization of text is very supportive in calculating whether lengthy document meets user need or not and is worth reading for supplementary information. In large text documents, text summarization software process summarizes the document in the time user reads the first paragraph. Basic of summarization process is to decrease the length and details of document while retaining its overall meaning and main points. Though it is hard to teach computer to analyze semantics and interpret meaning [6].

We as humans review text by reading the complete document for full understanding and then write summary by highlighting the main points [6]. But computers do not have language capabilities like humans so alternative methods are used. Automated summarization process is divided into three steps [7]:

- In *preprocessing step* a structured representation of original text is obtained.
- In *processing step* algorithm transforms the text structure to summary structure.
- In *generation step* the final summary is obtained from summary structure.

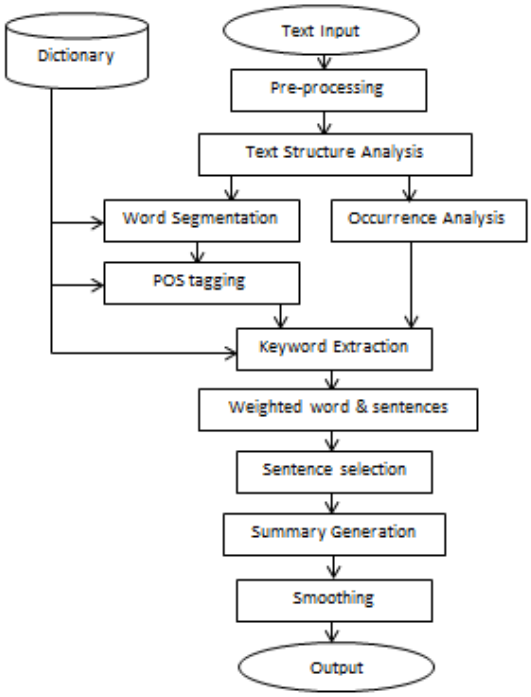

Figure 6: Text Summarization Process

5) Categorization

In this main themes of document are identified by placing document into predefined set of topics. Document is treated as "bag of words" by computer, actual information is not processed, only words that appear are counted and main topic that document covers is identified. It relies on thesaurus and relationships are identified by looking broad terms, narrow terms, synonyms etc. Tools designed for this purpose ranks the document on the basis of order in which document has most content on particular topic. Topic tracking and categorization can be combined to find relevance of document for

person finding information on a topic. Supervised learning algorithms can be used to learn classifiers from labeled documents and perform classification automatically on unlabeled documents. Flow diagram of text categorization is shown in figure 7. Taking example of set of labeled documents from a source *D=[d1,d2,……dn]*, which belongs to set of classes *C=[c1,c2……cn]*. Text categorization trains the classifiers using these documents and assigns categories to new documents. Training phase arranges *n* documents in *p* separate folders, where each folder corresponds to one class. After this training data set is prepared by using feature selection process.

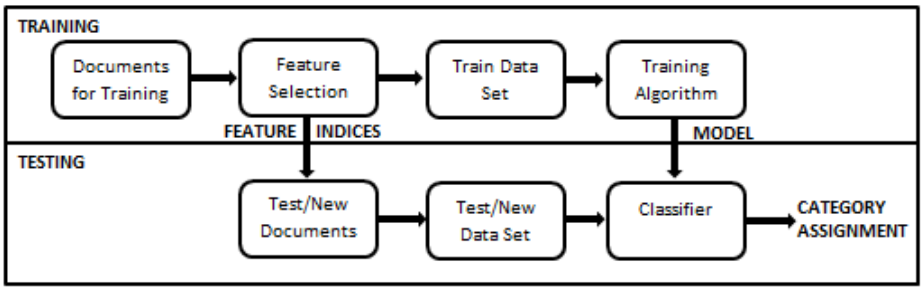

Figure 7: Categorization Process

6) Clustering

It is grouping of similar documents but different from categorization since documents are clustered on fly instead of use of predefined topics. Benefit in this is that documents can appear in multiple subtopics, ensuring that useful documents will not be omitted from search results. Clustering algorithm creates vector of topics for each document and measures the weight of how well document fits into each cluster.

Problem with statistical text clustering lies with high dimensionality of feature space. Standard clustering techniques are not efficient enough to deal with huge feature set.

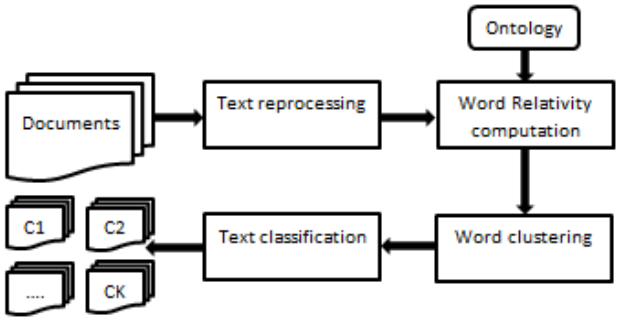

Figure 8: Clustering Process

7) Association Rule Mining

This technique is used to find relationships among large set of variables in data set. It has huge advantage in field of industry, it is discovering relationship among large set of variables, while database of records is present each containing two or more variables and their corresponding values. It checks frequently occurring combination of variable-value. In ARM a relationship can contain two or more variables. It is mostly used to find out which item customers buys together. In text mining ARM is used to study relationships among topics. Taking an example than when book *Data Mining Concepts and Techniques* is brought, 40% of the time the book *Database System* is brought together, and 25% of the time the book *Data Warehouse* is brought together. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc.

ARM finds 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 $\Rightarrow$ B = 1/5 = 0.2 = 20%
- Confidence of A $\Rightarrow$ B = $\frac{\text{Support of A} \Rightarrow \text{B}}{\text{Support of A}} = \frac{0.2}{0.4} = 0.5 = 50\%$
- Lift of A $\Rightarrow$ B = $\frac{0.2}{(0.4)(0.6)} = 0.83$

IV. CONCLUSION

Text mining is a growing technology which is also known as Text Data Mining (TDM) or Knowledge-Discovery in Text (KDT). Text mining is the process of extracting exciting and non-trivial information and knowledge from shapeless data. It is very young field of study of Data mining. Extracting meaningful information from unstructured data is not an easy task, and decision to choose correct method for text mining is also the decision of researcher. Automatic text mining still needs lots of research for finding efficient and easily implementable method.

REFERENCES

[1] Anmol Kumar, Amit Kumar Tyagi, Surendra Kumar Tyagi, "Data Mining: Various Issues and Challenges for Future A Short discussion on Data Mining issues for future work", www.ijetae.com (ISSN 2250-2459 (Online), Volume 4, Special Issue 1, February 2014) International Conference on Advanced Developments in Engineering and Technology (ICADET-14)

[2] Mr. Rahul Patel, Mr. Gaurav Sharma," A survey on text mining techniques", International Journal Of Engineering And Computer Science ISSN:2319-7242Volume 3 Issue 5 May, 2014 Page No. 5621-5625

[3] S. Jusoh and H. M. Alfawareh, "Agent-based knowledge mining architecture," in Proceedings of the 2009 International Conference on Computer Engineering and Applications, IACSIT. Manila, Phillipphines: World Academic Union, June 2009, pp. 602–606.

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

[5] Fred Popowich, "Using Text Mining and Natural Language Processing for Health Care Claims Processing",

[6] Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications".

[7] 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.

[8] 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.

[9] 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.

.

