International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-4 Issue-6 December-2012

Text mining: A Brief survey

Falguni N. Patel[1], Neha R. Soni[2]

Computer Engineering Department
Sardar Vallabhbhai Patel Institute of Technology
Vasad[1,2]
Email: svit_vasad.gen@rediffmail.com
ME. (Computer Engg.) Student, Associate Professor
E-mail:[1]falgunimanish19@gmail.com, [2]neha_ripal@yahoo.com

Abstract

The unstructured texts which contain massive amount of information cannot simply be used for further processing by computers. Therefore, specific processing methods and algorithms are required in order to extract useful patterns. The process of extracting interesting information and knowledge from unstructured text completed by using Text mining. In this paper, we have discussed text mining, as a recent and interesting field with the detail of steps involved in the overall process. We have also discussed different technologies that teach computers with natural language so that they may analyze, understand, and even generate text. In addition, we briefly discuss a number of successful applications of text mining which are used currently and in future.

Keywords

Classification, Clustering, Information Extraction, Question Answering, Summarization, Text Mining, Topic Tracking.

1. Introduction

Text mining is a technique which extracts information from unstructured data and find pattern which is novel and unknown earlier. It is also known as knowledge discovery from text (KDT), deals with the machine supported analysis of text [1].Text documents are in semi-structured or unstructured format datasets such as emails, full-text documents, HTML files etc. The problem of Knowledge Discovery from Text (KDT) is to extract explicit and implicit concepts and semantic relations between concepts. Its aim is to get insights into large quantities of text data.

Humans have the ability to distinguish and apply linguistic patterns to text and humans can easily overcome obstacles that computers cannot easily handle such as slang, spelling variations and contextual meaning. However, although our language capabilities allow us to understand unstructured data, we lack the computer's ability to process text in large volumes or at high speeds. Therefore, Text mining help to computer for analysis task on unstructured data. Text mining used for knowledge management and Human resource management, Customer Relationship Management, Technology Watch, Natural Language Processing and Multilingual Aspect.

In this paper, we described text mining as a whole process in second section. In third section we explained basic technologies of text mining. Last section of our paper discusses about the different application of text mining.

2. Text Mining Process

The steps involved in the overall process of the text mining are depicted in the Figure 1[2].

A. Text Preprocessing

The text pre-processing step further divided into number of sub steps as follows:

1) Tokenization :

Text document has a collection of sentences. This step divide whole statement into words by removing spaces, commas etc.

2) Stop word Removal:

This step involves removing of HTML,XML tags from web pages. Then process of removal of Stop

words like "a", "of" etc. is performed. Finally word stemming is applied.

3) Stemming:

These techniques are used to find out the root/stem of a word. Stemming converts words to their stems. E.g. Flying, Flew word to Fly. The algorithm proposed by Port, known as a Port's stemming algorithm is widely used for the same [3].

B. Text Transformation / Feature Generation

Text transformation means to convert text document into bag of words or Vector space document model notation, which can be used for further effective analysis task.

C. Feature Selection/Attribute Selection

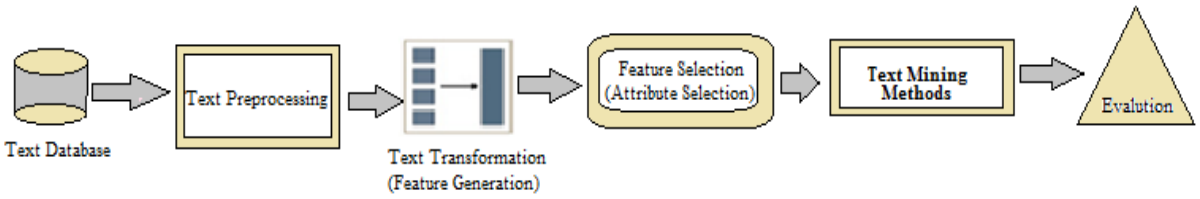

Fig- 1. General text mining process flow.

This phase mainly performs removing features that are considered irrelevant for mining purpose. This procedure give advantage of smaller dataset size, less computations and minimum search space required.

D. Text mining methods

There are different text mining methods as in Data mining had been proposed such as: Clustering, Classification, Information retrieval, Topic discovery, Summarization, Topic extraction.

E. Interpretation or Evaluation

This phase includes Evaluation and Interpretation of results in terms of calculating Precision and Recall, Accuracy, F measure etc.

3. Basic Technologies

Current Trend is that computer would be work as human for language analysis, understanding, Generation of text etc. Recent Technologies can achieve this using following methods. In this section we discuss all technologies with example so it will be useful to work further in the interested area.[4]

A. Information Extraction

Information extraction technique identifies key phrases and Information extraction technique identifies key phrases and relationship within a text [5]. For that it uses pattern matching method. Pattern matching means matching predefined sequences of text with user text. This technique is very useful in analysing large text dataset. The extracted information by IE cannot be represented directly into a structured form. Hence post processing is required [6].

B. Summarization

This technology is condensing the source text into a shorter version preserving its information. Human cannot manually summarize large documents[7].In big research organization, researcher do not have time to read all documents so they summarize document and highlight summary with main points. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. An abstractive summarization attempts to develop an understanding of the main concepts in a document and then express those concepts in natural language. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document. One of the strategy most widely used by text summarization tools, a sentence extraction extracts important sentences from an article by statistically weighting

the sentences. Further heuristics such as position Information are also used for summarization.
Figure- 2 describes the overall process of summarization.

C. Topic Tracking

Topic tracking is used to facilitate user by maintaining the topic searched or viewed by user previously. Next time system predict user's other search documents related to previous topic very effectively [8].Topic detection studies the problem of detecting new and upcoming topics in time ordered documents. The methods are frequently used in order to detect and monitor news tickers or news broadcasts.

D. Classification (Categorization)

Classification technique classify text documents into predefined class label (categories)[9].

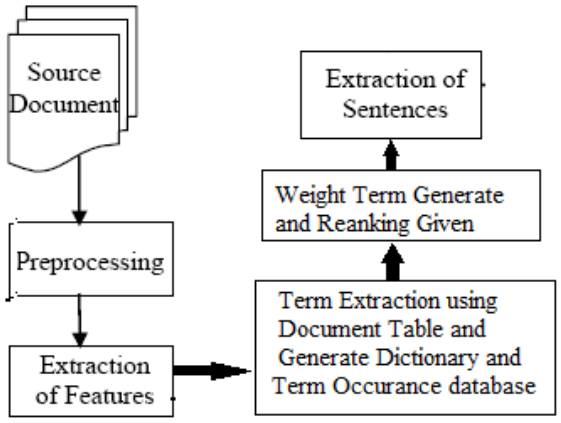

Fig-2 Summarization using Extracting Sentence

Classification has been used in many applications like Mobile sms classification [10], online customer feedback classification, business reports classification etc. Classification can be integrated with topic tracking to classify the documents by topic and thus making the process faster. Figure 3 depicts the process of classification.

E. Clustering

Clustering is a technique which has no predefined class labels but using similarity measures between different objects, it put most similar object in one class and dissimilar in another class. In figure 4 the general steps used in document clustering are described. Very first words are separated and then weights are applied to each of them. Then similarity is calculated and last different clustering algorithms like H, Partitioning, M, methods can be applied [11].

F. Concept Linkage

Concept linkage finds related documents who share common concepts. The primary goal of concept linkage is to provide browsing for information rather than searching for it as in information retrieval. For example in biomedical, concept link used to link diseases and treatment. In future, Text mining can be applied as a concept linkage to discover new treatments by associating treatments that had been used in related fields.

G. Information Visualization

To increase the use or acquisition of knowledge, we need interactive visual representation of abstract data.

The government can use information visualization to identify terrorist networks or to find information about crimes that may have been previously thought unconnected [12].It could provide possible relationships between suspicious activities so that they can investigate connections. This evolution can be

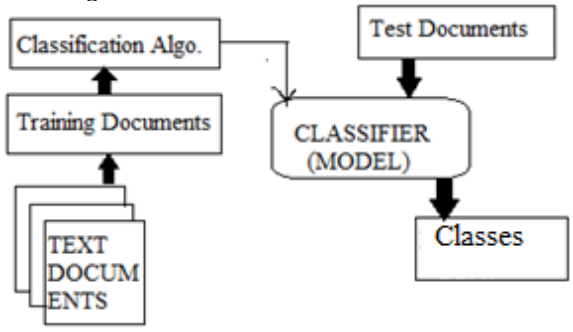

Fig-3 Classification model: Construction and Prediction

read in the invention of visual artifacts, from writing to mathematics, to maps, to diagrams, to visualizing. Visualization required generally three steps for Information visualize: Data Preparation, Data analysis and Extraction and Visualization mapping.so finally we get data space in visualization target as required by managers, marketers

H. Question Answering

Many websites that are equipped with question answering technology, allow end users to "ask" question to the computer and get exact or related answer [13]. Question and Answering technique utilizes multiple text mining methods for the same. Basic question and Answering shown in figure 5, is discussed below.

First step is the passage retrieval (PR) method. It allows passages with the highest probability of containing the answer to be retrieved, instead of simply recovering the passages sharing a subset of words with the question. Second step is Answer Extraction-aims to establish the best answer for a given question. It is based on a supervised machine-learning approach. It consists of two main modules, one for attribute extraction and the other for answer selection.

I. Association Rule Mining

Association rule mining (ARM) is a technique used to discover relationships among a large set of variables in a data set [14]. It has been applied to a variety of industry settings and disciplines but has not been widely used in the social sciences, especially in education, counselling, and associated disciplines.

Database containing two or more variables and their respective value, ARM determines variable value by calculating variable's frequency.

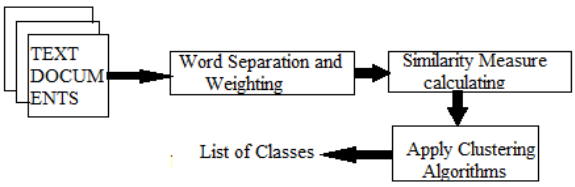

Fig-4 Basic Clustering Concept

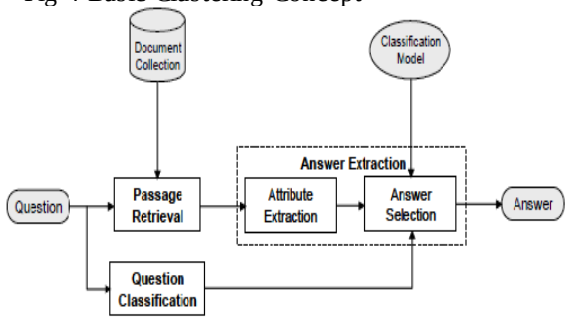

Fig -5 Process of Questions Answering [13]

ARM used in decision making process. It discover customer purchasing pattern and find relation or associations between different products. Therefore marketing concept is clear for organization to decide product selling approach.

J. Natural Language Processing (NLP)/ Computational Linguistics

The Goal of NLP is to design and build a computer system that will analyze, understand and generate NLP [15]. Application includes machine translation of one human language text to another human-language text, used in fiction, robotic systems etc. Thus it is useful for enabling the use of human language for providing a summary after understanding any text document, for commands and queries understanding and analysis purpose.

4. Application of Text Mining

Text mining has a very high commercial value. It is an emerging technology for analysing large collection of unstructured documents for the purpose of extracting interesting and non-trivial pattern or knowledge.

There are many domain specific application of Text mining, some of the applications we had explained here:

1) Customer Profile Analysis:

Companies use text mining to draw out the occurrences and instances of key terms in large blocks of text such as articles, Web pages, complaint forums [16]. The software converts the unstructured data formats into topic structures and semantic networks which are important data drilling tools. By studying the semantic network, one can learn the general tone of the complaints, reasons for complaining. It also finds common words used in complaints and their relationships to other words in the text via semantic weight [17].

2) Security applications

Many text mining software packages are marketed for security applications, especially monitoring and analysis of online plain text sources such as Internet news, blogs, etc. for national security purposes. It

also involved in the study of text encryption/decryption.

3) Biomedical Application

Text Mining is used in biomedical for identification and classification of technical terms in the domain of molecular biology corresponding to concepts.

4) Company Resource Planning

Mining company's reports and correspondences for activities, so its resource status and problems reported can be handled properly and future action planned can be design.

5) Open-ended survey responses

Analysing a certain set of words or terms that are commonly used by respondents to describe the pros and cons of product or service, suggesting common misconceptions or confusion regarding the items in the study. As per response of customers, industry takes the advantage of this for marketing [18].

6) Competitive Intelligence

Enabling companies to organize and modify the company strategies according to present market demands and the opportunities based on the information collected by the company about themselves, the market and their competitors, and to manage enormous amount of data for analysing to make plan [18].

7) Customer Relationship Management (CRM):

Rerouting specific requests automatically to the appropriate service or supplying immediate answers to the most frequently asked questions.

8) Technology Watch:

Identification of the relevant Science and Technology literatures and extraction of the required Information from these literatures efficiently, text mining techniques are used extensively [19].

9) Organize Repositories of document-related meta- information:

Automatic text categorization methods are used to create structured metadata, which is used for searching and retrieving relevant documents based on query [20].

10) Human Resource Management:

Mainly with applications aiming at analysing staff's opinions, monitoring the level of employee satisfaction as well as reading and storing CVs for the selection of new personnel TM is used. Often utilized to monitor the state of health of a company by means of the systematic analysis of informal documents.

5. Conclusion and Future work

Text mining, also known as Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. We also discuss the General Process of Text encoding and mining. Current text mining products and applications are designed for trained knowledge specialists. Future text mining tools, as part of the knowledge management systems, should be readily usable by technical users as well as management executives. Efforts are still required in developing systems that interpret natural language queries and automatically performs the appropriate mining operations. Text mining now used in security purpose like bug or roomer sms classifies on mobile station and removed. Therefore in context this mobile sms classification is also required more future work in this area of text mining.

Reference

[1] Haralampos Karanikas and Babis Theodoulidis Manchester, "Knowledge Discovery in Text and Text Mining Software", Centre for Research in Information Management (2001), UK.

[2] Munyaradzi Chiwara, Mahmoud Al-Ayyoub, Mohammad Sajjad Hussain,Rajan Gupta,"CSE-634 Data mining: Text Mining", ppt presentation.

[3] M. Porter. An algorithm for suffix Stripping and stemming. Program, pages 130–137, 1980.

[4] 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.

[5] 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.

[6] Sergio Bolasco, Alessio Canzonetti , Francesca Della,Ratta-Rinald and Bhupesh K. Singh, (2002), "Understanding Text Mining: a Pragmatic Approach", Roam, Italy.

[7] Vishal Gupta,Guruprit Lehal,"A survey of Text Summarization Extractive Techniques", Journal Of Emerging Technologies In Web Intelligence, Vol. 2, No. 3, August 2010.

[8] Sungjick Lee and Han-joon Kim, "News Keyword Extraction for Topic Tracking", 4th International conference on Networked Computing and Advanced Information Management, IEEE (2008), Korea. 554-559.

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

[10] Deepshikha Patel, Monika Bhatnagar,"Mobile sms Classification", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307 (Online), Volume-I, Issue-I, March 2011.

[11] Shady Shehata, Member, IEEE, Fakhri Karray, Senior Member, IEEE, and Mohamed S. Kamel, Fellow, IEEE," An Efficient Concept-Based Mining Model for Enhancing Text Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 10, October 2010.

[12] Chung Wong, Paul Whitney and Jim Thomas," Visualizing Association Rules for Text Mining", International Conference, Pacific Northwest National Laboratory, USA, 1-5.

[13] Antonio Juárez-González, Alberto Téllez-Valero, Claudia Delicia-Carral," Using Machine Learning and Text Mining in Question Answering", Language Technologies Group, Computer Science Department, National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico.2006.

[14] Dion H. Goh and Rebecca P. Ang," An introduction to association rule mining: An application in counselling and help seeking behaviour of adolescents", Journal of Behaviour Research Methods39 (2), Singapore, 259-266,2007.

[15] U. Ackermann, B. Angelini, F. Brugnara, M. Federico, D.Giuliani, R. Gretter, G. Lazzari and H. Niemann, "Speed Data: Multilingual Spoken Data Entry", International Conference, IEEE, Trento, Italy, 2211 -2214.

[16] Shantanu Godbole, and Shourya Roy, India, "Text to Intelligence: Building and Deploying a Text Mining Solution in the Services Industry for Customer Satisfaction Analysis", IEEE, 441-448, 2008.

[17] Kuan C. Chen, Purdue University Calumet, USA "Text Mining e-Complaints Data From e-Auction Store With Implications for Internet Marketing Research "Journal of Business & Economics Research – May, 2009 Volume 7, Number 5.

[18] Seth Grimes (2005), "The developing text mining market", white paper, Text Mining Summit Alta Plana Corporation**,** Boston, 1-12.

[19] Ronald Nell Kostoff, "Text Mining For Global Technology Watch", article, Office of Naval Research, Quincy St. Arlington, 1-27, Year 2003.

Ms.Falguni N. Patel. From Vadodara She is doing M.E in Computer Engineering (Second Year) from SVIT,Vasad, She has got B.E. in Information Technology Degree from SVIT,Gujarat University in 2005.From 2005 to 2011,She worked for SVIT as a Lecturer in I.T. Department. She published papers in International and National conferences.

Ms. Neha Ripal Soni from Vadodara. She completed her M.E. in Computer Science with first rank and is a gold medallist from D.D.I.T. University, Gujarat, India and pursuing Ph.D. in Data Mining from CHARUSAT University, Changa.She is presently working as an Associate Professor at SVIT, Gujarat Technological University. She has published a number of papers in the proceedings of National and International level conferences. She is a life member of Computer Society of India (CSI) and ISTE.