ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 9, September 2014

A Concise Survey on Text Data Mining

Patil Monali S[1], Kankal Sandip S[2]

M.Tech Scholar, Department of Computer Science & Engineering, Maharashtra Institute of Technology (MIT), Aurangabad, Maharashtra, India [1]

Assistant Professor, Department of Computer Science & Engineering, Maharashtra Institute of Technology (MIT), Aurangabad, Maharashtra, India [2]

Abstract: In recent days the use of internet is growing rapidly. The data used and shared by the users on internet is in huge amount which is available in unstructured, semi- structured and structured form such as images, texts, audios or videos. For analysis and processing of such immense data, data mining came into picture. Data mining is the process of retrieving previously unknown and significant information from given set of data. Among the data available in digital form over the internet, 85% of data available is in unstructured form. Most of the data used is in text form such as Electronic mail, Internet chat, World Wide Web, Digital libraries, Electronic Publications, and Technical reports etc. For the purpose of knowledge discovery and information retrieval from such textual data text mining is used. Text mining is a kind of data mining technique responsible for retrieving valuable information from collection of text.

In this paper, focus is on concept of text mining, text mining process flow, data mining methods used in text mining such as Clustering, Topic detection, Information Extraction and Natural Language Processing. Also presenting some real world applications of text mining.

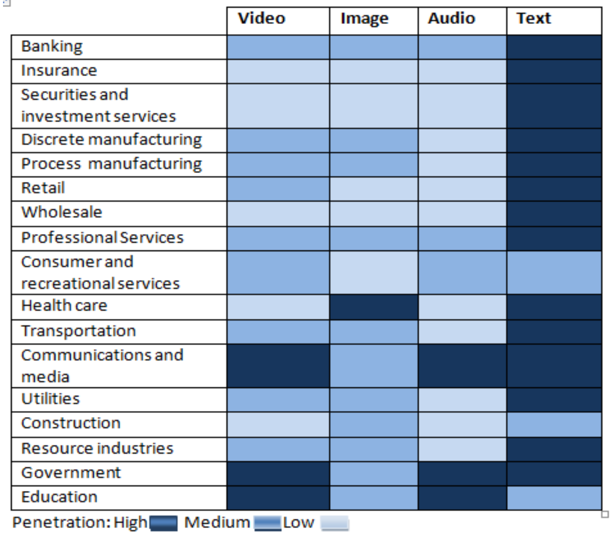
Keywords: Unstructured Data, Semi-structured Data, Structured Data, Data mining, Natural Language Processing (NLP), Information Extraction (IE).

I. INTRODUCTION

An immense amount of data is figured out in textual form. Near about 80% of information in text form can be generated through web documents, customer follow-ups, emails, weblog articles, news, social networking sites like face book, twitter and also by industrial data[1]. Nowadays lots of the data present in organization, businesses and industries are stored electronically, in the form of text databases [2]. With respect to survey of HP 70%, Gartner 80%, Jerry Hill (Teradata) of data is unstructured data [3]. According to example in [4] there is 30 billon pieces of content shared on face book every month in text form, 235 terabytes data collected by US library of congress by April 2011. Such a Big data consisting of unstructured, semi-structured data which is continuously increasing in vast amounts so that Analysis, Processing and Organization of such textual data is crucial part for business organization. Although mankind is capable of handling operations on text file like grammatical correction but lacking in analyzing, processing and organizing textual information.

Natural language processing has numerous troubles in processing text, Text mining is solution for this juncture. Text mining is the process of furnishing and extracting information from such unstructured data.

Fig. 1 shows that, maximum amount of data used by various sectors is in textual form. Text mining is one of the important fields of data mining dealing with unstructured or semi-structure data. Text mining introduces enumerate model from linked research domain like Statistics, Machine learning, Information Retrieval (IR), Information Extraction (IE), Natural language processing (NLP) etc. Numerical measures can be deriving by enforcing Text analysis methods to unstructured textual information.



	Video	Image	Audio	Text
Banking	▨	▨	▨	■
Insurance	□	□	□	■
Securities and investment services	□	□	□	■
Discrete manufacturing	▨	▨	□	■
Process manufacturing	▨	▨	□	■
Retail	▨	□	□	■
Wholesale	□	□	□	■
Professional Services	▨	▨	▨	■
Consumer and recreational services	▨	□	▨	▨
Health care	□	■	□	■
Transportation	▨	▨	□	■
Communications and media	■	▨	■	■
Utilities	▨	▨	□	■
Construction	□	▨	□	▨
Resource industries	▨	▨	□	■
Government	■	▨	■	■
Education	■	▨	■	▨

Penetration: High ■ Medium ▨ Low □

Fig. 1. Types of data available and generated by various sectors

II. TEXT MINING DEFINITIVE

Computer system endured the analytical thinking of text. By addressing different research areas we can define Text mining in number of ways which deals by view of each area.

- Information Retrieval (IR): One way that assumes text mining relation correlated to Information Retrieval i.e. retrieval of concept from text.
- Knowledge Organization Structure (KOS) for examining knowledge organization structure,

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

examining the patterns of gigantic amount of text data for the goal of processing and capturing information in Knowledge Organization Structure (KOS).

- Text mining is the process of computation of extracting information from bulk quantity of data with the help of representing subordinate data to robust form.
- Text mining is the process of uncovering new unidentified information by retrieving entropy from numerous written and digital resources.

III. TEXT MINING PROCESS AND METHODS

A) Process of Text Mining

1) Document Gathering:
In the first step, the text documents are collected which are present in different formats[1]. The document might be in form of pdf, word, html doc, css etc.

2) Document Pre- Processing:
In this process, the given input document is processed for removing redundancies, inconsistencies, separate words, stemming and documents are prepared for next step, the stages performed are as follows [1][2]:

a) Tokenization:
The given document is considered as a string and identifying single word in document i.e. the given document string is divided into one unit or token[1].

b) Removal of Stop word:
In this step the removal of usual words like a, an, but, and, of, the etc. is done [6].

c) Stemming:
A stem is a natural group of words with equal (or very similar) meaning. This method describes the base of particular word. Inflectional and derivational stemming are two types of method[4]. One of the popular algorithm for stemming is porter's algorithm[5]. e.g. if a document pertains word like resignation, resigned, resigns then it will be consider as resign after applying stemming method[6].

3) Text Transformation:
A text document is collection of words (feature) and their occurrences. There are two important ways for representations of such documents are Vector Space Model and Bag of words [7].

4) Feature Selection (attribute selection):
This method results in giving low database space, minimal search technique by taking out irrelevant feature from input document. There are two methods in feature selection i.e. filtering and wrapping methods.

5) Data mining/Pattern Selection:
In this stage the conventional data mining process combines with text mining process. Structured database uses classic data mining technique that resulted from previous stage [7].

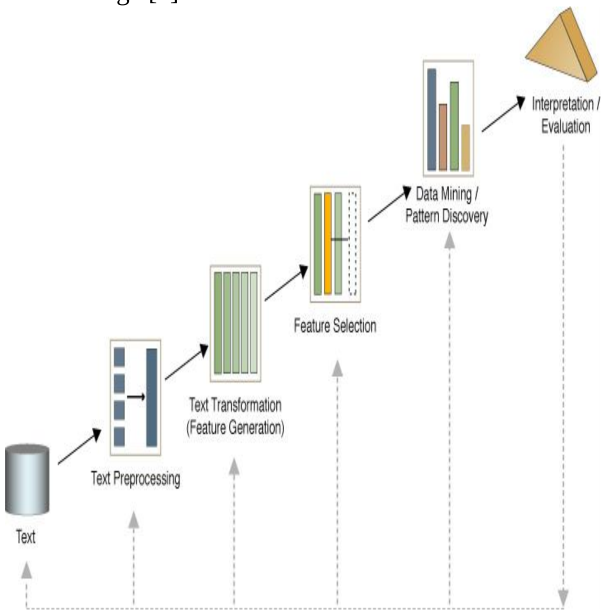

Text
Text Preprocessing
Text Transformation (Feature Generation)
Feature Selection
Data Mining / Pattern Discovery
Interpretation / Evaluation

Fig. 2. Text Mining Process flow.

6) Evaluate:
This stage Measures the outcome. This resulted outcome can be put away or can be used for next set of sequence [7].

B) Data Mining Methods for Text Mining**:**

1) Information Extraction:
The bulk of data is available in an unstructured form. The drawback of conventional system is that it considers that the information to be mined is in the form of relational databases. Information Extraction system addresses this problem of traditional system. The system looks for text stream available in the given document and determines the relationship and similarity between words and phrases.
As illustrated in fig. 2 The database drawn from given document by Information Extraction system is supplied to knowledge discovery databases for advanced mining processes. After mining knowledge from extracted data, Information Extraction system can predict information missed by the previous extraction using discovered rules DISCOTEX [9].

2) Topic Detection:
Conventional keyword search engines are restricted to a given data model and cannot easily adapt to unstructured, semi-structured or structured data [10]. Topic detection system overcomes this issue.
Topic is an originative or event directly related along with all events and activities. Topic tracking mechanism helps to detect such an event which is related to target topic [11]. Topic Detection and Tracking (TDT) refers to automatic techniques for finding topically related material in streams of data (e.g., newswire and broadcast news). Work on TDT began about a year ago, is now expanding, and will be a regular feature at future Broadcast News workshops [11].

Topic Detection system uses two Topic Mining Models which are as follows:

1. Vector Space Model:
Despite of its simple data structure without using any explicit semantic information, the vector space model enables very efficient analysis of huge document collections. It was originally introduced for indexing and information retrieval [12] merely applies to document extraction scheme and also in text data mining technique. This model forms the matrices of high dimension for respective document.
The compounding of document and term is nothing but text document[13] . There are many such documents that appears in the form of structured and semi-structured format. Text file like book, spreadsheets, telephone directory etc. Terms are words or group of word that are retrieved from document.
Vector model represent each document as vector and each value in vector is represent word number at that value represent the number of occurrences of the word in document. Suppose collection of document $M=(m_1,m_2,m_3,\ldots\ldots.,m_N)$ and collection of terms $N=(n_1,n_2,n_3,\ldots,n_L)$ then we need to model vector $V_{i\,j} = (v_{i,1}, v_{i,2},\ldots\ldots..,v_{i,L})$in L dimension space. In Boolean expression if $V_{i,j}=1$, then term n_i belongs to m_i and if $V_{i,j}=0$ then n_i does not belongs to m_i .

2. Probabilistic Topic Model:
Machine learning researchers have developed probabilistic topic modelling, a suite of algorithms that aim to discover and annotate large archives of documents with thematic information [14].

3) Natural Language Processing:
Text Mining is widely used in field of Natural Language Processing and Multilingual aspects. In NLP, Text Mining applications are also quite frequent and they are characterized by multilinguals [8]. Use of Text Mining techniques to identify and analyse web pages published in different languages, is one of its example [8]. The General Architecture for Text Engineering (GATE) is a framework for the development and deployment of language processing technology in large scale (Cunningham, Maynard, Bontcheva, & Tablan, 2002). GATE can be used to process documents in different formats including plain text, HTML, XML, RTF, and SGML.[EMERGING TECHOF TMT]

4) Clustering:
Text mining becomes more challenging because of its characteristics such as volume, dimensionality, scarcity and complex semantics involved in it. These characteristics require clustering techniques to be scalable to large and high dimensional data, and be able to handle sparsity and semantics. Typical text clustering activity involves with the document representation, document similarity measure and clustering techniques. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity) [18].

IV. APPLICATIONS

1) Web Search Enhancement:
In text mining, by using *text categorization* techniques such as CatS [14]. The presentation of result is by sorting them into (a hierarchy of) clusters which may be displayed to the user in a variety of ways, e.g. as a separate expandable tree (vivisimo.com) or arcs which connect Web pages within graphically rendered “maps” (kartoo.com) [18].

2) Bioinformatics:
Nowadays, biomedical articles have occupied major area in different applications [17]. The aim of text mining in bioinformatics is to allow researchers to explore advance knowledge in the field of biomedical in an efficient way.

3) Business Intelligence:
Text mining techniques helps for determining particular topic or event as in business decision support system amount of cutting down the cost of predicting future work is an important task [18].

4) Open-Ended Survey Responses*:*
Analyzing 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].

5) Human Resource Management:
For purpose of recruiting the candidate by reading and writing their CVs text mining is used [19]. Also for growth of particular organization for the application prefers the monitoring of level of customer, examining staff’s satisfaction text mining techniques applied.

6) Security Application:
In network security for encryption and decryption techniques, text mining methodologies are used with the intention of security at national level application, the analysis and monitoring of documents like plain texts, emails, web blog articles text mining is used [19].

V. CONCLUSION

In this paper, we studied most dynamic field of Text in area of Data mining, the text mining processing flow, methodologies of Data mining used in Text mining, the application areas of text mining such as Bioinformatics, Business Intelligence, and Human Resource Management, Security application, Open ended survey responses and web search Enhancement etc. Text mining provides solution for applications or areas where extracting, processing and analysis of text from data warehouses is crucial part.

ACKNOWLEDGEMENT

I am genuinely thankful to Prof. Mr. Kankal Sandip S., who directed me in right way for penning above work, of Computer science and technology Department, MIT, Aurangabad.

REFERENCES

[1] Vandana Korde and C. Namrata Mahender, ”Text Classification and Classifiers :A Survey”, International Journal of Artificial Intelligence & Application, Vol.3, No.2, March 2012.

[2] R. Sagayam, S. Srinivasan and S. Roshni”, A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques,” International Journal Of Computational Engineering Research Vol. 2 Issue.

[3] http://www.abbottanalytics.com

[4] http://www.planetdata.eu/sites/default/files/presentations/Big_Data_Tutorial_part4.pdf

[5] Andreas Hotho, Andreas Nurnberger and Gerhard PaaB, “A Brief Survey of Text Mining”, May 13, 2005

[6] http://www.isical.ac.in/~acmsc/TMW2014/M_mitra.pdf

[7] Lokesh Kumar and Parul Kalra Bhatia,”Text Mining:Concept,Process,Applications,” Journal of Global Research in Computer Science Volume 4, No. 3, March 2013 .

[8] Vishal Gupta and 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] N. Kanya and S. Geetha, “Information Extraction: A Text Mining Approach”,International Conference on Information and Communication Technology in Electrical Sciences, IEEE (2007).

[10] Guoliang Li, Beng Chin Ooi, Jianhua Feng ,Jianyong Wang and Lizhu Zhou, ” EASE: An Effective 3-in-1 Keyword Search Method for Unstructured, Semi-structured and Structured Data”.

[11] *Charles L. Wayne*, ”Topic Detection & Tracking (TDT) Overview & Perspective”.

[12] G.Salton, A. Wong, and C. S. Yang.” A vector space model for automatic Indexing”,*Communications of the ACM*, 18(11):613–620, 1975.

[13] Sushmita Mitra, Tinku Acharya “ Data Mining Multimedia, Soft Computing, and Bioinformatics”.

[14] Jonathan G. Fiscus and George R. Doddington, ” Topic Detection and Tracking Evaluation Overview”.[15] Charu C Aggrawal and Chengxiang Zhai,”Mining Text Data”.

[15] David M Blei, Princeton University, ”Introduction to Probabilistic Topic Models”.

[16] Shaidah Jusoh 1 and Hejab M. Alfawareh,” Techniques, Applications and Challenging Issue in Text Mining”, IJCSI International Journal

[17] Of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012.

[18] Novi Sad J. Math, Milos Radovanovic and Mirjana Ivanovic,” Text Mining: Approaches and Applications” Vol. 38, No. 3, 2008, 227-234.

[19] Falguni N. Patel, Neha R. Soni, “Text mining: A Brief Survey”.

[20] Seth Grimes, “The developing text mining market”, white paper, Text Mining Summit Alta Plana Corporation, Boston, 1-12, 2005.

BIOGRAPHY

Patil Monali S, Pursuing M. Tech from MIT College of engineering, Aurangabad. She is assisting as a lecturer in Deogiri Institute of Technology and Management Study, Aurangabad. She has accomplished B.E (Information Technology) from M.I.T, Aurangabad. She also completed Diploma in Information Technology from Government Polytechnic, Aurangabad. Her domain of pursuit is Data Mining as well Big Data, Text mining, Natural Language Processing.