Volume 5, Issue 2, February 2020 ISSN (online): 2456-0006
International Journal of Science Technology Management and Research
Available online at: www.ijstmr.com

Text Mining Techniques for Similarity Measure Between Documents - A Recent Survey

Gitanjali Gupta
P.G. Research Scholar
C.S.E. Department
JIT Borawan Khargone

Mr. Kapil Shah
Asst. Professor
C.S.E. Department
JIT Borawan Khargone

Abstract:- ***Nowadays most of the information in business, industry, government and other institutions is stored in text form into database and this text database contains semi structured data. A document may contain some largely unstructured text components like abstract additionally few structured fields as title, name of authors, date of publication, category, and so on. Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. Text mining tool would retrieve a particular document and pre-process it by checking format and character sets. Then document would go through a text analysis phase. Text analysis is semantic analysis to derive high quality information from text. Many text analysis techniques are available; depending on goal of organization combinations of techniques could be used. Sometimes text analysis techniques are repeated until information is extracted. In this paper we proposed a recent survey on various text mining techniques for similarity measure***

Keywords:- ***Document, Text, Similarity, Pre-processing, terms***

I. INTRODUCTION

Text mining is a new field that attempts to collect meaningful information from natural language text. It may be loosely characterized as the process of analyzing text to extract information that is useful for particular purposes. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with algorithmically. Text is the most common vehicle for the formal exchange of information. The field of text mining usually deals with texts whose function is the communication of factual information or opinions, and the motivation for trying to extract information from such text automatically is compelling even if success is only partial. The phrase "text mining" is generally used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful information.

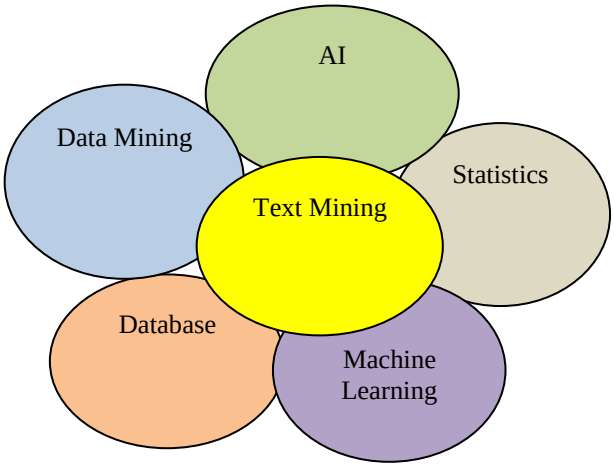

Figure 1 Text mining interaction with other fields

Text mining process performs the following steps

1. Collecting unstructured data from different sources Text mining interaction with other fields available in different file formats such as plain text, web pages, pdf files etc.
2. Pre-processing and cleansing operations are performed to detect and remove anomalies. Cleansing process makes sure to capture the real essence of text available and is performed to remove stop words stemming and indexing the data.
3. Processing and controlling operations are applied to audit and further clean the data set by automatic processing.
4. Pattern analysis is implemented by Management Information System (MIS).

5. Information processed in the above steps are used to extract valuable and relevant information for effective and timely decision making and trend analysis.

TEXT MINING ISSUE

Text mining has following issues

1. Large numbers of small documents vs. small numbers of large documents.
2. Excluding certain characters, short words, numbers, etc. Excluding numbers, certain characters, or sequences of characters, or words
3. Include lists, exclude lists (stop-words). Specific list of words to be indexed can be defined.
4. Synonyms and phrases. Synonyms, such as "sick" or "ill", or words that are used in particular phrases where they denote unique meaning can be combined for indexing.
5. Stemming algorithms. An important pre-processing step before indexing of input documents begins is the stemming of words.

II. LITERATURE SURVEY

In 2010 Russell Albright, Richard Foley proposed "Listening to the Twitter Conversation**".** They discuss the use of text mining, visualization, and the HTTP Procedure to provide a complete understanding of the Twitter conversation. They showed that SAS provides great tools to understand Twitter, from collecting the data with SAS procedures to analyzing the data with SAS. Text Analytics provides the ability to analyze the Twitter conversation in its entirety, filter out the noise, and understand the topics of the conversation. Understanding individuals who focus on topics provides more clarity to what is relevant to your brand. In order to interpret the data, they view the data in a way that is easy to understand. New graphics being developed by SAS, and other standard SAS graphs, allow for quick and easy visualization of what is happening on the Twitter conversation. With SAS, a complete set of products to access, analyze, and visualize the Twitter conversation is available. [1].

In 2011 James Thomas, John McKnight proposed "Applications of text mining within systematic reviews". They describe the application of four text mining technologies, namely, automatic term recognition; document clustering, classification and summarization, which support the identification of relevant studies in systematic reviews. They showed that text mining technologies to improve reviewing efficiency are considered and their strengths and weaknesses explored. They conclude that these technologies do have the potential to assist at various stages of the review process. They are relatively unknown in the systematic reviewing community, and substantial evaluation and methods development are required before their possible impact can be fully assessed. [2].

In 2012 Su Gon Cho and Seoung Bum Kim proposed "Identification of Research Patterns and Trends through Text Mining". They crawled the keywords from the abstracts in IIE Transactions, one of the representative journals in the field of Industrial Engineering from 1969 to 2011. They applied a low-dimensional embedding method, clustering analysis, association rule, and social network analysis to find meaningful associative patterns of the keywords frequently appeared. They revealed research trends and patterns of the Industrial Engineering field from one of the representative journals in Industrial Engineering by using text mining. They employed the dimensional reduction method, clustering analysis, and social network analysis to draw out meaningful patterns of They stimulate further investigation in applying appropriate text and data mining tools to various applications in both academia and industry [3].

In 2013 K. L. Sumathy M. Chidambaram proposed "Text Mining Concepts, Applications, Tools and Issues An Overview". They give an overview of concepts, applications, issues and tools used for text mining. Due to the rapid growth of digital data made available in recent year's knowledge discovery and data mining have attracted great attention with a forthcoming need for turning data into useful information and knowledge. Consequently there is growing research interest in the topic of text mining. In general text mining consists of analyzing large amount of text documents by extracting key phrases; concepts etc., and prepare the text processed for further analysis with data mining techniques [4].

In 2014 Sonali Vijay Gaikwad proposed "Text Mining Methods and Techniques". They presented survey and discuss such successful techniques and methods to give effectiveness over information retrieval in text mining. These types of situations where each technology may be useful in order to help users are also discussed. They presented overview techniques methods and challenging issue in text mining. They addressed the most challenging issue in developing text mining systems. Four methods of text mining term based, phrase based, and concept based and pattern taxonomy model discussed. To provide improved understandable information for mining the documents Graphical Visualization is used [5].

In 2015 E. Alan Calvillo, Alejandro Padilla proposed "Searching Research Papers Using Clustering and Text Mining". They proposed a better classification of research papers, the architecture works with a database of knowledge related with the topics of programming, databases and operating systems. They evaluates a way to optimize the information to be located within a structured framework with an initial knowledge base that helps the easy categorization of information by implementing a clustering for fast. They shows as future work is consider to extend the search engine into another kind of devices using Android or IOS that's to generate a portable application to make searches in different kind of devices[7].

In 2015 Nadir Zanini and Vikas Dhawan proposed "Text Mining: An introduction to theory and some Applications". They briefly described the techniques of TM and some of its applications. TM has a variety of potential applications in the field of education. To date, these applications have been carried out by teachers and assessment experts without using advanced techniques such as TM, but TM

allows the possibility of implementing these applications on a more comprehensive scale. The availability of novel data could lead, at least in principle, to novel measurement and research designs to address old and new research questions [8].

In 2016 Ramzan Talib, Muhammad Kashif Hanif proposed "Text Mining Techniques that affect the accuracy and relevance of results are identified. They present a brief overview of text mining techniques that help to improve the text mining process. Domain knowledge integration, varying concepts granularity, multilingual text refinement, and natural language processing ambiguity are major issues and challenges that arise during text mining process. In future research work, they will focus to design algorithms which will help to resolve issues presented in this work [9].

In 2016 Abhishek Kaushik and Sudhanshu Naithani proposed "A Comprehensive Study of Text Mining Approach". They proposed a review of text mining techniques, tools and various applications Text mining is one of the fastest growing fields today. From last few years text mining (sentiment analysis) is largely being used to predict the results of elections at national and state level which is most significant development in the field recently. Text mining is also being used by industry and it is generating the sheer amount of knowledge which cannot even consume by humans. They tried to present an overview of text mining approach with its techniques, tools and applications [10].

In 2016 R. Janani, Dr. S. Vijayarani proposed "Text Mining Research: A Survey". They discussed about the text mining techniques and its applications. Text mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources. They delineated the various text mining techniques such as Information Extraction, Information retrieval, Natural Language processing, Categorization and Clustering. They also defined text mining processing flow, applications of text mining and issues in text mining. Mining text in different languages may be a major problem, since text mining tools and techniques ought to be able to work with several languages and multilingual languages. [10].

In 2017 Binling Nie and Shouqian Sun proposed "Using Text Mining Techniques to Identify Research Trends: A Case Study of Design Research". They give information about scientific literature in design research They presents a bibliometric, network-theoretic and text-based analysis of the design research area during the last 12-year period (2004–2015). They used first in-depth study on keeping track of the current advances in the design research area by using text mining techniques. Furthermore, the result shows that the developed methods are universal and could be applied to manage the knowledge of various research fields. [11].

In 2018 Said A. Salloum, Mostafa Al-Emran proposed "Using Text Mining Techniques for Extracting Information from Research Articles" They collected, and textually analyzed through various text mining techniques, three hundred refereed journal articles in the field of mobile learning from six scientific databases, namely: Springer, Wiley, Science Direct, SAGE, IEEE, and Cambridge. They present study demonstrates a comprehensive overview about text mining and its current research status. According to the surveyed literature, there is a limitation in discussing the issue of information extraction from research articles using data mining techniques. They perceive that information extraction and data mining techniques were never applied to the mobile learning field. [14]

III. TEXT MINING TECHNIQUES

Text mining techniques can be understood at the processes that go into mining the text and discovering insights from it. These text mining techniques generally employ different text mining tools and applications for their execution. Some of text mining techniques are

1. Information Extraction

This is the most famous text mining technique. Information exchange refers to the process of extracting meaningful information from vast chunks of textual data. This text mining technique focuses on identifying the extraction of entities, attributes, and their relationships from semi-structured or unstructured texts. Whatever information is extracted is then stored in a database for future access and retrieval. The efficacy and relevancy of the outcomes are checked and evaluated using precision and recall processes.

2. Information Retrieval

Information Retrieval (IR) refers to the process of extracting relevant and associated patterns based on a specific set of words or phrases. In this text mining technique, IR systems make use of different algorithms to track and monitor user behaviors and discover relevant data accordingly. Google and Yahoo search engines are the two most renowned IR systems.

3. Categorization

This is one of those text mining techniques that is a form of "supervised" learning where in normal language texts are assigned to a predefined set of topics depending upon their content. Thus, categorization or rather Natural Language Processing (NLP) is a process of gathering text documents and processing and analyzing them to uncover the right topics or indexes for each document. The co-referencing method is commonly used as a part of NLP to extract relevant synonyms and abbreviations from textual data.

4. Clustering

Clustering is one of the most crucial text mining techniques. It seeks to identify intrinsic structures in textual information and organize them into relevant subgroups or 'clusters' for further analysis. A significant challenge in the clustering process is to form meaningful clusters from the unlabeled textual data without having any prior information on them. Cluster analysis is a standard text mining tool that assists in data distribution or acts as a pre-processing step for other text mining algorithms running on detected clusters.

5. *Summarization*

Text summarization refers to the process of automatically generating a compressed version of a specific text that holds valuable information for the end-user. The aim of this text mining technique is to browse through multiple text sources to craft summaries of texts containing a considerable proportion of information in a concise format, keeping the overall meaning and intent of the original documents

essentially the same. Text summarization integrates and combines the various methods that employ text categorization like decision trees, neural networks, regression models, and swarm intelligence.

IV. COMPARISON

We compare these techniques on the basis of process used in the techniques

S. No.	Techniques	
1	Information Extraction	Extracting meaningful information from vast chunks of textual data
2	Information Retrieval	Process of extracting relevant and associated patterns based on a specific set of words or phrases.
3	Categorization	Process of gathering text documents and processing and analyzing them to uncover the right topics or indexes
4	Clustering	Identify intrinsic structures in textual information and organize them into relevant subgroups or 'clusters' for further analysis.
5	Summarization	Text summarization refers to the process of automatically generating a compressed version of a specific text that holds valuable information for the end-user

REFERENCE

1. James Thomas, John McNaught and Sophia Ananiadoub Applications of text mining within systematic reviews Received 2 September 2010, Accepted 28 Published online 11 April 2011 in Wiley Online Library(wileyonlinelibrary.com) DOI: 10.1002/jrsm.27.
2. Su Gon Cho and Seoung Bum Kim Identification of Research Patterns and Trends through Text Mining International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012 Manuscript received March 20, 2012; revised April 12, 2012.The authors are with the School of Industrial Management Engineering, Korea University, Seoul, Korea (e-mail: sbkim1@korea.ac.kr.)
3. K. L. Sumathy, M. Chidambaram, Text Mining: Concepts, Applications, Tools and Issues An Overview International Journal of Computer Applications (0975 – 8887) Volume 80 No.4, October 2013.
4. Sonali Vijay Gaikwad Archana Chaugule Text Mining Methods and Techniques International Journal of Computer Applications (0975 – 8887) Volume 85 – No 17, January 2014.
5. E. Alan Calvillo, Alejandro Padilla1,Jaime Muñoz1 Searching Research Papers Using Clustering and Text Mining All content following this page was uploaded by Julio Cesar Ponce on 12 November 2015.
6. Ramzan Talib, Muhammad Kashif Hanify, Shaeela Ayeshaz, and Fakeeha Fatima Text Mining: Techniques, Applications and Issues (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7 No. 11, 2016.
7. D. Jasmine Guna Sundari A Study of Various Text Mining Techniques International Journal of Advanced Networking & Applications (IJANA) Volume: 08, Issue: 05 Pages: 82-85 (2017) Special Issue.
8. Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem and Khaled Shaalan Using Text Mining Techniques for Extracting Information from Research Articles © Springer International Publishing AG 2018 K. Shaalan et al. (eds.), Intelligent Natural Language Processing: Trends and Applications, Studies in Computational Intelligence 740, https://doi.org/10.1007/978-3-319-67056-0_18.
9. Latinka Todoranova, Bonimir Penchev, Radka Nacheva Using Text Mining To Classify Research Papers 17 th International Multidisciplinary Scientific Geo Conference SGEM 2017.
10. Morgan Kaufmann Data Mining: Concepts and Techniques Series in Data Management Systems Morgan Kaufmann Publishers, July 2011. ISBN 978-0123814791 .
11. Binling Nie and Shouqian Sun Using Text Mining Techniques to Identify Research Trends: A Case Study of Design Research Institute of Industrial Design, College of Computer Science, Zhejiang University, Hangzhou 310027, China; ssq@zju.edu.cn Appl. Sci. 2017
12. Said A. Salloum, Mostafa et al "Using Text Mining Techniques for Extracting Information from Research Articles" Springer International Publishing AG 2018 K. Shaalanet al. (eds.), Intelligent Natural Language Processing: Trends and Applications, Studies in Computational Intelligence 740, https://doi.org/10.1007/978-3-319-560