International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Text Mining: Survey on Techniques and Applications

M. Uma Maheswari[1], Dr. J. G. R. Sathiaseelan[2]

[1]Research Scholar, Department of Computer Science, Bishop Heber College, Trichy 620017

[2]Associate Professor and Head, Department of Computer Science, Bishop Heber College, Trichy 620017

Abstract: ***Text Mining has turned into a vital research zone. Text Mining discloses new and already obscure data, by consequently removing data from various composed assets. In this paper, a Survey of Text Mining strategies and applications have been exhibited. The unstructured text documents from various source contains huge amount of information which are not to be used for any processing to extract useful information. Text mining is the process of extracting significant information or knowledge or patterns from the available unstructured text documents.Text mining tasks includes text categorization, text clustering, document summarization and sentiment analysis.There exist different techniques and tools to mine the text and discover valuable information for future prediction and decision making process. This paper discussed general idea of text mining, explains various techniques used to extract useful information, discussed number of text mining application and tools used for text mining process. In addition of that, this paper discussed issues in the field of text mining.***

Keywords: Text Mining, Knowledge Discovery, Classification, Clustering, Summarization

1. Introduction

Today the internet has massive amount of text in the form of digital libraries, repositories, and other textual information such as blogs, reports, reviews, news, social media network and e-mails. It is difficult task to find out appropriate patterns and trends to extract important knowledge from this large volume of data [1]. However, large amounts of information such as textual information are unstructured, and defy simple attempts to make sense of it. Manual analysis of this unstructured textual information is increasingly impractical, and as a result, text mining techniques are being developed to mechanize the process of analyzing this information.

Text mining is the process of discovering new, previously unknown information, knowledge through automated extraction of information from often large amounts of unstructured text. In general, the unstructured text is easy for people, but very complex for computer program. In particular, difficulties with automated text comprehension are caused by the fact that the human/ natural language:

- Ambiguous terms and phrases
- Often strongly relies on the context and background knowledge for defining and conveying meaning.
- Strongly based on commonsense knowledge and reasoning
- Is influenced by and is influencing people's mutual interactions.

To overcome these difficulties, the text mining employs a set of algorithms for converting text into structured data and the quantitative methods used to analyze these data. The fundamental objective of text mining is to enable users to extract data from text based resources and manages the operations like retrieval, extraction, summarization, categorization and clustering.

Text mining techniques are continuously applied in industry, academia, web applications, internet and other fields [2]. Application areas like search engines, customer relationship management system, filter emails, product suggestion analysis, fraud detection, and social media analytics use text mining for opinion mining, feature extraction, sentiment, predictive, and trend analysis [3].

Extraction of valuable information from a corpus of different document is a tedious and tiresome task. The selection of appropriate technique for mining text reduces the time and effort to find the relevant patterns for analysis and decision making. The objective of this paper is to analyze different text mining techniques which help to perform text analytics effectively and efficiently from large amount of data. Moreover, the issues that arise during text mining process are identified. Text Mining is the disclosure of new obscure data, by consequently separating data from various documents. A key component is the connecting together of the separated data together to shape new actualities or new theories to be investigated by more ordinary methods for experimentation. Text mining is not quite the same as what are comfortable within web search. In pursuit, the client is regularly searching for something that is as of now known and has been composed by another person. The issue is pushing aside all the material that right now is not important to the requirements keeping in mind of the end goal to locate the applicable data. In text mining, the objective is to find obscure data, which something that nobody yet knows thus couldn't have however recorded.

This paper is categorized as follows: The section II explains process of text mining, section III discuss different techniques of text mining, application of text mining techniques are presented in section IV. Section V provides the conclusion of this work.

2. Process of Text Mining

Text mining is the process of converting unstructured data to structured data for extracting useful information. Figure 1 shows the text mining process.

There are five steps under text mining process: Data Collection, Data Pre-processing, and Data Transformation, Data Analyse and Result evaluation.

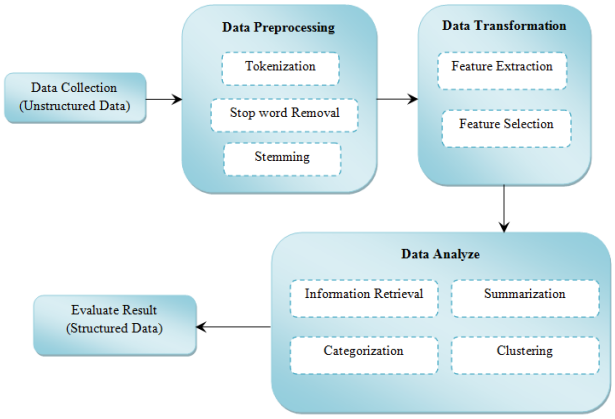

Figure 1: Text Mining Process

2.1. Data Collection

In this step the unstructured data was collected from various sources and it can be in the form of reports, blog, reviews and news.

2.2. Data Pre-processing

In this step the collected data was preprocessed for removing redundancies, inconsistencies, separate words and stemming. In the tokenization, the data was divided into single word i.e. token.

The data contains unwanted words like, a an, the, but, and, of, etc. These words are called as stop words. Stop words are removed in this step.

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. One of the popular algorithms for stemming is porter's algorithm. e.g. if a document pertains word like resignation, resigned, resigns then it will be consider as resign after applying stemming method.

2.3. Data Transformation

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

In feature extraction, the useful meaning words are extraction from the document.In feature selection, relevant words are selected. There are two methods in feature selection i.e. filtering and wrapping methods.

2.4. Data Analyse

The processed data was analyzed using text mining methods such as information retrieval, categorization, classification and summarization.

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

2.5. Evaluation

This step evaluates the results in terms of precision, recall and accuracy.

3. Text Mining Techniques

This section explains text mining techniques.

3.1. Clustering

Clustering is an unsupervised process to classify the text documents in groups by applying different clustering algorithms. In a cluster, similar terms or patterns are grouped extracted from various documents. Clustering is performed in top-down and bottom up manner. In NLP, various types of mining tools and techniques are applied for the analysis on unstructured text. Different techniques of clustering are hierarchical, distribution, density, centroid, and k-mean [4]

Zhang et al.[5] used cosine to calculate a correlation similarity between two projected documents in a low-dimensional semantic space and performed document clustering in the correlation similarity measure space.

Hassan [6] proposed an algorithmic framework for partition clustering of documents that maximizes the sum of the discrimination information provided by documents. It exploits the semantic that term discrimination information provides to yield clusters that are describable by their highly discriminating terms.

Qimin et al [7] proposed a feature cluster-based vector space model (FC-VSM) which used the text feature clusters co-occurrence matrix to represent document and proposed to identify non-contiguous phrases in the text preprocessing stage. This method improves the quality of text clustering.

Wei et al. [8] exploited an ontology hierarchical structure for word sense disambiguation to assess similarity ofwords. The experiment results showed better clustering performance for ontology-based methods considering the semantic relations between words.

3.2. Categorization

Text categorization (TC – also known as text classification, or topic spotting) is the task of automatically sorting a set of documents into categories (or classes, or topics) from a predefined set [9].

Using supervised learning algorithms, the objective is to learn classifiers from known examples (labeled documents) and perform the classification automatically on unknown examples (unlabeled documents). Figure.2 shows the overall flow diagram of the text categorization task. Consider a set of labeled documents from a source D = [d1,d2,….dn] belonging to a set of classes C = [c1,c2,…,cp]. The text categorization task is to train the classifier using these documents, and assign categories to new documents. In the training phase, the n documents are arranged in p separate folders, where each folder corresponds to one class. In the next step, the training data set is prepared via a feature selection process.

Text data typically consists of strings of characters, which are transformed into a representation suitable for learning. It is observed from previous research that words work well as features for many text categorization tasks.

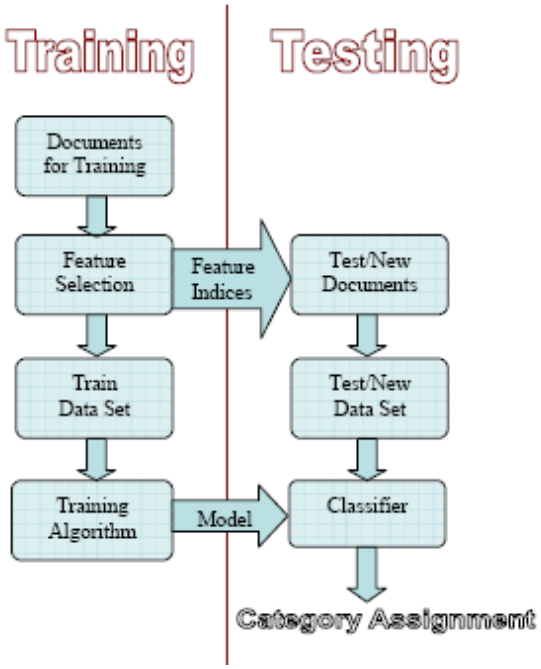

Figure 2: Flow Diagram of Text Categorization

In the feature space representation, the sequences of characters of text documents are represented as sequence of words. Feature selection involves tokenizing the text, indexing and feature space reduction. Text can be tokenized using term frequency (TF), inverse document frequency (IDF), term frequency inverse document frequency (TFIDF) or using binary representation. Using these representations the global feature space is determined from entire training document collection.

In text manifold categorization method, the text documents are treated as vectors in an n-dimensional space, where every dimension corresponds to a term.

Then the metrics such as the cosine of the angle between two documents can be defined. However this space may be intrinsically located on the low dimensional manifold. The metric therefore should be defined according to the properties of manifold so as to improve the text categorization furthermore. The whole process is illustrated as Figure 3.

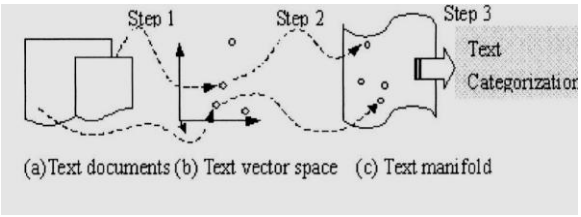

Figure 3: Framework of text categorization on text manifold

Text categorization are automatically assigned to appointed species according to text content. Similar texts are assigned to the same species through calculating the similarity among the texts. After the process of pattern aggregation for the word matrix, the numbers of words are greatly decreased, then TF.IDF method is applied to constructing the VSM.

As the dimensions of the text are greatly decreased through the process of the P-L, the method decreases the learning time, and advances the speed and the of text categorization

Huang [10] introduced text categorization technique called VSM_WN_TM which is a combination of Vector Space Model (VSM), WordNet ontology, and Probabilistic Latent Semantic Analysis (PLSA) topic modeling. It also used the support vector machine for classification purposes.

Zheng [11] proposes a novel approach for text categorization based on a regularization extreme learning machine (RELM) in which its weights can be obtained analytically, and a bias-variance trade-off could be achieved by adding a regularization term into the linear system of single-hidden layer feed forward neural networks.

Tang [12]presents a Bayesian classification approach for automatic text categorization using class-specific features.

3.3. Summarization

Text summarization is a process of collecting and producing concise representation of original text documents [13]. In past automatic text summarization was performed on the basis of occurrence a certain word or phrase in document. Later on, additional methods of text mining were introduced with standard text mining process to improve the relevance and accuracy of results [14]

Ferreira et al [15] proposed a new summarization system thateasily combines different sentence scoring methods in

orderto obtain the best summaries depending on the context. Pal et al. [16] has proposed the Wordnet based method to identify the semantics behind various input text by making use of Lesk algorithm.

Animesh Ramesh et al. [17], introduced SentenceRank algorithm which uses statistical and semantic analysis between sentences for computing importance of sentences in order to summarize text. This algorithm build semantic graph where nodes are considered as sentences and edges are semantic relatedness between that sentences which is calculated by with the help of WordNet. Rank these nodes using a ranking algorithm and select top ranked sentence as summary

Alguliev et al. [18] designed an evolutionary optimization algorithm for multi-document summarization. This algorithm creates a summary by collecting the salient sentence from the multiple documents. This approach utilizes the summary to document collection, sentence-to-document collection, and sentence-to-sentence collection to choose the most important sentences from the multiple documents.

Xiong and Lu [19] introduced an approach for multi-document summarization using Latent Semantic Analysis (LSA). Among the existing multi-document summarization approaches, the LSA was a unique concept, which uses the latent semantic information rather than the original features. It has chosen the sentence individually to remove the redundant sentences.

3.4. Sentiment Analysis

Sentiment analysis is a type of natural language processing for tracking the mood of the public about a particular product or topic. Sentiment analysis, which is also called opinion mining, involves in building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets.

Lin et al [20] proposes a novel probabilistic modeling framework called joint sentiment-topic (JST) model based on latent Dirichlet allocation (LDA), which detects sentiment and topic simultaneously from text. A re-parameterized version of the JST model called Reverse-JST, by reversing the sequence of sentiment and topic generation in the modeling process. The topics and topic sentiments detected by JST are indeed coherent and informative.

Arun et al [21] proposed review analyzer for analyzing consumer product reviews from review collections. It is based on performing the sentimental words' analysis for sentiment classification. Cui, A. et al. [22] showed that sentiment analysis of tweets is a challenging task due to multilingual and informal messages. The paper tackles this problem by analysis of emotion tokens. Emotion is themood of a person depicted from the words in the tweet. Emotion can be sad, happy, angry, etc. The proposed approach has two steps. First, emotion tokens are extracted from the message. Second, graph propagation algorithm plots the tokens at different polarities. Finally, sentiment analysis algorithm analyses and classifies these emotion tokens.

4. Application of Text Mining

4.1. Email Spam Filtering

E-mail is an effective, fast and reasonably cheap way to communicate, but it comes with a dark side: spam. Today**,** spam is a major issue for internet service providers, increasing their costs for service management and hardware\software updating; for users, spam is an entry point for viruses and impacts productivity**.** Text mining techniques can be implemented to improve the effectiveness of statistical-based filtering methods**.**

4.2. Social media data analysis

Today, social media is one of the most prolific sources of unstructured data; organizations have taken notice. Social media is increasingly being recognized as a valuable source of market and customer intelligence, and companies are using it to analyze or predict customer needs and understand the perception of their brand. In both needs Text analytics can address by both analyzing large volumes of unstructured data, extracting opinions, emotions and sentiment and their relations with brands and products.

4.3. Business Intelligence

Text mining plays a significant role in business intelligence that helps organizations and enterprises to analyze their customers and competitors to take better decisions. It provides a deeper insight about business and gives information how to improve the customer satisfaction and gain competitive advantages. The text mining tools like IBM text analytics, Rapid miner, and GATE help to take decisions about the organization that generate alerts about good and bad performance, market changeover that help to take remedial actions. It also helps in telecommunication industry, business and commerce applications and customer chain management system

5. Conclusion

The availability of huge volume of text based data need to be examined to extract valuable information. Text mining techniques are used to analyze the interesting and relevant information effectively and efficiently from large amount of unstructured data. In this survey of text mining, several text mining techniques and its applications in various fields have been discussed. Text mining algorithms will give us useful and structured data which can reduces time and cost. Hidden information in social network sites, bioinformatics and internet security etc. are identified using text mining is a major challenge in these fields. The advancement of web technologies

has lead to a tremendous interest in the classification of text documents containing links or other information.

References

[1] N.Padhy,P. Mishra , and R.Panigrahi, (2012), "The Survey of Data Mining Applications And Feature Scope" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012

[2] S.-H. Liao, P.-H. Chu, and P.-Y.Hsiao (2012), "Data mining techniques and applications–a decade review from 2000 to 2011," Expert Systems with Applications, vol. 39, no. 12, pp. 11 303–11 311, 2012

[3] W. He, (2013) "Examining students online interaction in a live video streaming environment using data mining and text mining," Computers in Human Behavior , vol. 29, no. 1, pp. 90–102, 2013

[4] B. L. Narayana and S. P. Kumar, (2015) "A new clustering technique on text in sentence for text mining," IJSEAT , vol. 3, no. 3, pp. 69–71, 2015

[5] T. Zhang, Y. Y. Tang, B. Fang, and Y. Xiang, (2012) "Document clustering in correlation similarity measure space," IEEE Trans. Knowl. Data Eng. , vol. 24, no. 6, pp. 1002–1013, Jun. 2012

[6] M.T. Hassan (2015), "Document Clustering by Discrimination Information Maximization", InformationSciences,Volume 316,87-106, 2015

[7] C. Qimin, G. Qiao,W. Yongliang, andW. Xianghua (2015), "Text clustering using VSM with feature clusters," Neural Computing and Applications, vol. 26, no. 4, pp. 995–1003, 2015

[8] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, (2015) "A semantic approach for text clustering usingWordNet and lexical chains," Expert Systems with Applications, vol. 42, no. 4, pp. 2264–2275, 2015.

[9] F. Sebastiani (2005), "Text categorization", Alessandro Zanasi (ed.) Text Mining and its Applications, WIT Press, Southampton, UK, pp. 109-129, 2005.

[10] Huang, Yinghao, Xipeng Wang, and Yi Lu Murphey. (2014) "Text categorization using topic model and ontology networks." Proceedings of the International Conference on Data Mining (DMIN). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2014.

[11] Zheng, W. B., Qian, Y. T., & Lu, H. J. (2013). Text categorization based on regularization extreme learning machine.Neural Computing &Applications ,22(3–4),447–456.

[12] B. Tang, H. He, P. M. Baggenstoss, and S. Kay (2016), "A Bayesian classification approach using class-specific features for text categorization," IEEE Trans. Knowl.Data Eng. , vol. 28, no. 6, pp. 1602–1606, Jun. 2016.

[13] B. A. Mukhedkar, D. Sakhare, and R. Kumar (2016), "Pragmatic analysis based document summarization," International Journal of Computer Science and Information Security , vol. 14, no. 4, p. 145, 2016

[14] C. P. Chen and C.-Y. Zhang, (2014) "Data-intensive applications, challenges,techniques and technologies: A survey on big data,"Information Sciences, vol. 275, pp. 314–347, 2014.

[15] Ferreira, R., Freitas, F., Cabral, L. d., Lins, R. D., Lima, R., Franc¸a, G., Favaro, L. (2014), "A Context Based Text Summarization System", 11th IAPR International Workshop on Document Analysis Systems. IEEE. 2014

[16] Pal, A.R., Saha, D. (2014) An approach to automatic text summarization using WordNet. In: Advance Computing Conference (IACC), 2014 IEEE International, pp. 1169, 1173 (2014). doi:10.1109/IAdCC.2014.6779492

[17] Animesh Ramesh, SrinivasaK.G,Pramod N (2014),"SENTENCERANK - a graph based approach to summarize text", Fifth International Conference on Applications of Digital Information and Web Technologies (ICADIWT), 2014.

[18] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, (2013) "Multiple documents summarization based on evolutionary optimization algorithm," Expert Systems with Applications, vol. 40, no. 5, pp. 1675–1689, 2013.

[19] S. Xiong and Y. Luo (2014), "A New Approach for Multi-document Summarization Based on Latent Semantic Analysis," Computational Intelligence and Design (ISCID), 2014 Seventh International Symposium on, Hangzhou, 2014, pp. 177-180.

[20] C. Lin, Y. He, R. Everson, S. Ruger, (2012) Weakly supervised joint sentiment-topic detection from text, IEEE Trans. Knowl Data Eng. 24, no. 6, 1134-1145

[21] M. ArunMonicka Raja, S. Godfrey Winster, S. Swamynathan (2012), "Review Analyzer: Analyzing Consumer Product Reviews from Review Collections", IEEE International Conference on Recent Advances in Computing and Software Systems, pp. 287-292, April 2012.

[22] A. Cui, M. Zhang, Y. Liu, S. Ma, (2011) Emotion Tokens: Bridging the Gap among Multilingual Twitter Sentiment Analysis, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 238–249.