Text Mining, Process and Models: An Overview

Mrs.Aswanandini.R,M.Sc.,M.Phil.,(Ph.D)
Assistant professor,Dept of Computer Science
KG College of Arts and Science,Coimbatore.

Abstract- *Text Mining is the process of extracting or understanding information from a set of texts. It is the use of automated methods for understanding the knowledge available in the text documents. It is also the process of analyzing text to extract information that is useful for a specific purpose. It can also work with semi-structured or unstructured data sets such as emails, text documents and HTML files etc. In this paper, the process and the models of text mining are discussed.*

Keywords- Text Mining, topic modeling, probabilistic modeling

I. INTRODUCTION

Text Mining field deals with tremendous amount of text data, which are created in a variety of forms such as social networks, E-mail, web articles, blog entries. Text data is a good example of unstructured information, which is one of the simplest forms of data that can be generated in most.

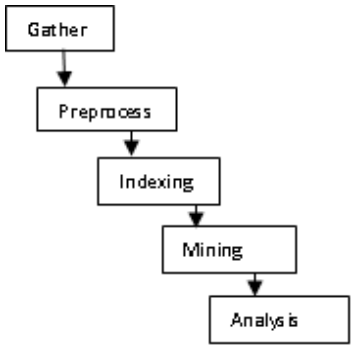
Unstructured text can be easily processed and perceived by humans, but is significantly harder for machines to understand. As a result, there is a need to design process , methods , models in order to effectively process the unstructured text in a wide variety of applications. Various process and modeling methods are used. In this paper, we have discussed about the process of text mining, topic modeling, the methods used in topic modeling, probabilistic modeling and its methods.

II. LITERATURE REVIEW

The fundamental process and techniques used in text domain are discussed [1]. Process involved in text mining to structure the unstructured data and the use of natural language processing in preprocessing [4] are used widely in this modern era.Topic Modeling is an efficient way to analyze the big unclassified text. The author gives a high level view of the methods used in topic modeling[12]. Recently, Probabilistic Classifiers have gained a lot of popularity and have shown to perform remarkably well. The simplest and the widely used classifier is the Naive Bayes classifier[1]. When compared to other methods in Probablistic Modeling, the Hidden Markov Model produces high level of accuracy[15]

III. TEXT MINING PROCESS

To mine the information efficiently, text mining involves a series of activities to be performed.



Gather: Text Mining starts with a collection of documents. This step involves the help of a search engine to find out the collection of text also known as corpus of texts which might need some conversion. These texts should be brought together in a format which will be helpful for the users to understand. Usually XML is the standard for text mining.

Pre-process: The dataset obtained is an unstructured dataset of documents which are pre-processed. The complexity of data pre-processing depends on the data sources used. It is a process of discovering sequential patterns in e-documents. This step allows the system to perform grammatical analysis of a sentence to read the text. It also analyzes the text in structures using Natural Language Processing. Preprocessing techniques are applied on the target data set to reduce the size of the data set.

Indexing: The next crucial process is indexing. In this process, the indexed representations collects a set of indexes and the information is expressed in natural language in the texts with the minimum loss of semantics.

Mining: At this point the text mining process merges with the traditional Data Mining process. It extract information, find patterns, and organize contents. This blends natural language processing, machine learning and semi-automated coding tools.

Analysis: Text mining consists of the analysis of text documents by extracting key phrases, concepts and prepares the text processed for further analyses with data mining techniques. This also allows users to combine the output of unstructured, text-based analytics with structured data to perform predictive analytics and data mining. This act as a tool for analysis and visualization of document collection.

III. MODELS USED IN TEXT MINING

The models used in text mining are topic modeling, probabilistic modeling and few methods under these modeling are discussed.

Topic Modeling:

Topic modeling is to discover patterns and how to connect documents that share similar patterns. To find natural groups of items, it is a method for unsupervised classification of documents, similar to clustering on numeric data. The documents are a mixture of topics, and the topic is a probability distribution over words. In other word, topic model is a generative model for documents. It specifies a simple probabilistic procedure by which documents can be generated.

Topic modeling has become a widely used tool for document management as a result of its superior performance. Massive amounts of information are dealt each day. As more and more data becomes accessible, it becomes difficult to access what we are searching for. So, we need tools and techniques to prepare, search and perceive huge quantities of information. Topic modeling provides us with strategies to organize, understand and summarize massive collections of textual data. It helps to find out hidden topical patterns that are present within the collection, illustrating documents consistent with these topics, using these illustrations to arrange, search and summarize texts.

The four methods that topic modeling rely on are Latent Dirichlet allocation (LDA), Text rank, Latent Semantic Analysis (LSA).

Latent Dirichlet allocation (LDA):

Latent Dirichlet allocation (LDA) is a popular method in a topic model. It treats each document as a mixture of topics that are present in the corpus and each topic as a mixture of words. Instead of being separated in to discrete groups, this allows the documents to overlap in terms of content that makes a typical use of natural language. The model proposes that each word in the document is attributable to one of the document's topics.

For example, consider the following set of documents as the corpus:

Suppose we have the following set of sentences:

- I like to have ice cream and coffee.
- I had coffee for breakfast.
- Dogs and Cats are petanimals.
- My sister adopted a kitten yesterday.
- Look at this children having icecream.

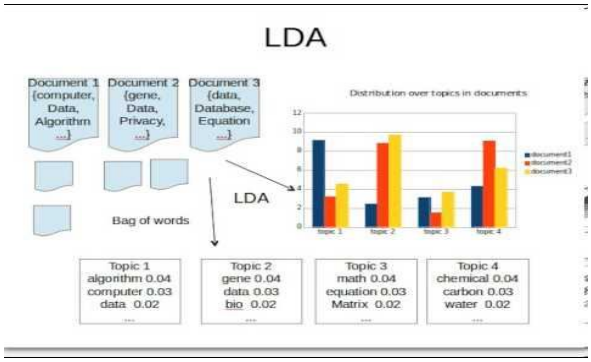
Latent Dirichlet allocation is a way that automatically discover topics that these sentences contain. For example, given these sentences and asked for 2 topics, LDA might produce something like

- sentence 1 and sentence 2: 100% Topic A
- sentence 3 and sentence 4: 100% Topic B
- sentence 5: 60% Topic A, 40% Topic B
- topic A: 30% icecream, 15% coffee, 10% breakfast
- topic B: 20% dogs, 20% cats, 20% petanimals, 15% children

In more detail, LDA represents documents as mixtures of topics that spit out words with certain probabilities. It assumes that documents are produced in the following fashion: when writing each document

- We should decide on the number of words N the document will have.
- Choose a topic mixture for the document over a fixed set of K topics.
- Generate each word w_i in the document by:
    - First picking a topic according to the multinomial distribution that we sampled above; for example, you might pick the food topic with 1/3 probability and the pet animals topic with 2/3 probability.
    - Using the topic to generate the word itself by multinomial distribution.

Assuming this generative model for a collection of documents, LDA then tries to backtrack from the documents to find a set of topics that are likely to have generated the collection.



Text Rank

Text Rank is a graph-based method that computes the importance of sentences. Sentences are regarded as nodes in the graph, while similarities among the sentences are regarded as edges between these nodes. Text Rank is similar to Page Rank. Text Rank can be used in order to find the most relevant sentences in text.

In order to find the most relevant sentences in text, a graph is constructed where the vertices of the graph represent each sentence in a document and the edges between sentences which are based on content overlap, by calculating the number of words that 2 sentences have in common. Based on this network of sentences, the sentences are fed into the Page rank algorithm which identifies the most important sentences. Now if a summary of text is to be extracted it takes only the most important sentences.

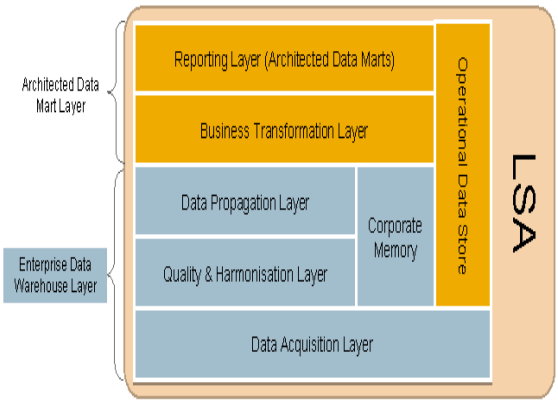
When node B is connected with node A, this means that node B has voted for node A. Meanwhile, the vote is represented by the similarity between the nodes. The more similar node B is to node A, the more important the vote is from node B to node A. Additionally a node with a higher score will give a more authoritative vote. When a node gets many votes, this means the node is very important and will have a higher score. Conversely, Page Rank just analyzes hyperlinks between web pages - a page is either connected with another page or not; but in our method the edge represents the similarity between nodes. Page Rank is improved by Text Rank by adding a weight to the edge, namely the Text Rank weighted graph model. In the Text Rank model, the importance of a node is related to the number of votes it obtains, the importance of nodes voting it and similarity between them.

Latent Semantic Analysis (LSA):

The competence of LSA's reflection of human knowledge has been recognized established in a variety of ways.The data acquisition layer takes the data from the source and distributes it in the BW system. This layer allows you fill all targets independently of each other, and even at different times. The data propagation layer is used for the applications. This should happen as quickly as possible, which is why you have the option of semantic partitioning in this layer.

The data propagation layer allows you to consolidate data and therefore can contain multiple levels. The Corporate Memory is filled independently of the update into the architected data marts which contains the complete history of the loaded data. This is used as a source for reconstructions in spite of the need to access the sources again.

However, LSA as currently practiced has some additional limitations. It manages to extract correct reflections of word and passage meanings quite well without these aids, but it must still be suspected of incompleteness or likely error on some occasions.



Probablistic modeling :

A probabilistic model is a model that uses probability theory to model the uncertainty in the data. A probabilistic model describes a set of possible probability distributions for a set of observed data, and the goal is to use the observed data to learn the distribution in the probabilistic model that can best describe the current data.

Naive Bayesian classifier:

The Naive Bayesian classifier is based on Bayes' theorem which makes independent assumptions between the predictors. Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. They are probabilistic, that calculate the probability of each category for a given sample, and the highest one is the

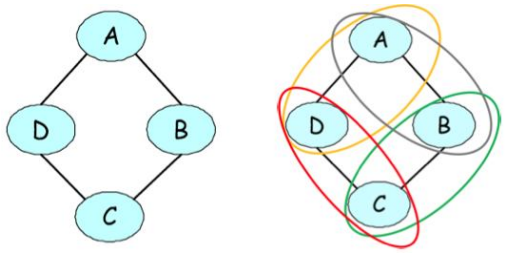
output category . The way they get these probabilities is by using Bayes' Theorem, which describes the probability of a feature, based on prior knowledge of conditions that might be related to that feature. Inspite of its simplicity, the Naive Bayesian classifier is the best method which is widely used because it often outperforms more sophisticated classification methods.

Hidden Markov models:

This model is a finite set of states which is related to a probability distribution. Transitions among the states are governed by a set of probabilities called *transition probabilities.* According to the associated probability distribution an outcome or *observation* can be generated.
Dynamic programming algorithms known as Viterbi algorithm is used for the first and the second problem and the Forward-Backward algorithm, respectively. The last one can be solved by an iterative Expectation-Maximization (EM) algorithm, known as the Baum-Welch algorithm.

Markov random fields:

Markov random fields is n-dimensional random process defined on a discrete lattice. The variables are connected by an edge if they directly influence each other. Markov random fields are useful for domains which can be termed as "soft constraints" between variables. Markov random fields can be characterized in terms of factorization of the joint distribution or conditional independence properties.



As an example, suppose that we are modeling voting preferences among persons *A*,*B*,*C*,*D*. Let's say that (*A*,*B*), (*B*,*C*), (*C*,*D*), and (*D*,*A*) are friends, and friends tend to have similar voting preferences. These influences can be naturally represented by an undirected graph.

IV. CONCLUSION

Text Mining is the process of extracting useful information from the data. It provides interesting patterns from large set of databases. This paper has given some information about topic modeling and probabilistic modeling and various methods used under these modeling. Today's growing interaction of text mining to some other fields, especially with machine learning, visualization and natural language processing, it is possible to design more effective and useful text mining system.

REFERENCES

[1] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. " A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques", In Proceedings of KDD Bigdas, Halifax, Canada, August 2017.

[2] P.Selvi, "An Analysis on removal of duplicate records using different types of Data Mining Techniques :A Survey", International Journal of Computer Science and Mobile Computing, November 2017.

[3] K.N.S.S.V.Prasad, S.K.Saritha, Dixa Saxena "A Survey Paper on Concept Mining in Text Documents" International Journal of Computer Applications (0975 – 8887) Volume 166 –No.11, May 2017.

[4] Sinoara R., Antunes J. & Rezende S. J Braz Comput Soc (2017) 23: 9. https://doi.org/10.1186/s13173-017-0058-7

[5] Abhishek Kaushik and Sudhanshu Naithani, "A Comprehensive Study of Text MiningApproach" International Journal of Computer Science and Network Security, VOL.16 No.2, February2016

[6] Ravindra Changala, Dr.D Rajeswara Rao "A Survey on Development of Pattern evolving Model for Discovery Of Patterns In Text Mining Using Data Mining Techniques "Journal of Theoretical and Applied Information Technology, 31st August 2017. Vol.95. No.16

[7] C.Uma, S.Krithika, C.Kalaivani "A Survey Paper on Text Mining Techniques", International Journal of Engineering Trends and Technology (IJETT), V40(4),225-229 , October 2016.

[8] Ramzan Talib , Muhammad Kashif Hanif, Shaeela Ay sha, and Fakeeha Fatima "Text Mining: Techniques, Applications and Issues", International Journal of Advanced Computer Science and Applications, Vol. 7 No. 11, 2016.

[9] Anne Kao, Steve R. Poteet "Natural Language Processing and Text Mining".

[10] Yogapreethi.n , Maheswari.s "A review on text mining in data mining", International journal on soft computing vol.7, No. 2/3, August 2016

[11] Arjun Srinivas Nayak, Ananthu P Kanive, Naveen Chandavekar , Dr. Balasubramani R," Survey on Pre - Processing Techniques for Text Mining", International Journal Of Engineering And Computer Science ISSN:

2319 -7242 Volume 5 Issues 6 June 2016,Page No. 16875-16879

[12] Ramzan Talib, Muhammad Kashif Hanif, Shaeela Ayesha, and Fakeeha Fatima "Text Mining: Techniques, Applications and Issues" International Journal of Advanced Computer Science and Applications,Vol. 7 No. 11, 2016

[13] Rubayyi Alghamdi, Khalid Alfalqi "A Survey of Topic Modeling in Text Mining", International Journal of Advanced Computer Science and Applications, Vol.6, No.1, 2015

[14] Ashok Srivastava, Mehran Sahami "Text Mining: Classification, Clustering, and Applications".

[15] Nihar Ranjan,Abhishek Gupta, Ishwari Dhumale, Payal Gogawale and Rugved Gramopadhye "A survey on text analytics and classification techniques for text documents" International Journal of Development Research Vol. 5, Issue, 11, pp. 5952 -5955, November,2015.

[16] I.Berin Jeba Jingle, Dr. J.Jeya a.Celin "MarkovModel for Discovering Knowledge in Text Documents" Journal of Theoretical and Applied Information Technology, 31st December 2014. vol.70 no.3.

[17] Lokesh Kumar , Parul Kalra Bhatia "Text mining: Concepts, process and applications "Journal of Global Research in Computer Science, Volume 4, No. 3 , March 2013.