A Survey on Text Mining, its Techniques and Applications

Shweta Gupta[#1], Kirti Joshi[#2]
[#]*CSE Dept, RIMT-IET, Mandi Gobindgarh, India*
[1]shweta5422@yahoo.co.in
[2] kirtijoshi11@gmail.com

Abstract: The process of Text mining is used to extract knowledge from large text documents. Text mining is becoming an important research field because it is necessity to extract knowledge from the huge volume of text documents available especially on the Web. Text documents can be in any form like e-mails, books, journal papers, E-libraries, news articles and so on. Although data mining and text mining seems to be similar terms but they are different at many issues, mainly, on the type of data both can handle. This review paper explains what text mining is and how data mining and text mining can be distinguished from each other. It also explains how text mining works and what are the different techniques used to implement text mining and areas where it can be applied.

Keywords: **Text Mining, Data Mining, Classification, Clustering.**

I. INTRODUCTION

Text mining is a process of extracting knowledge from large text documents. Text mining is also referred to as text data mining. Text mining is a growing new field that attempts to gather meaningful information from natural language text. Compared with the type of data stored in databases, text is unstructured, vague, and difficult to handle algorithmically. Text mining is a system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful information. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output. Typical text mining tasks include text categorization, text clustering, concept / entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling. A typical application is to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate a database or search index with the information extracted. The diagrammatic general process of Text Mining is shown here. While large-scale information technology has been evolving separate transaction and analytical systems, text mining provides the link between the two. Text mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software available are statistical, machine learning, and neural networks. Generally, any of four types of relationships are used which are given below:

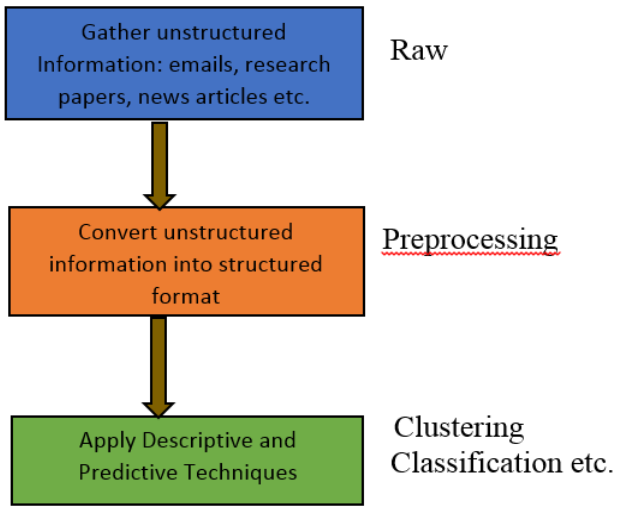

Fig. 1 General process of Text Mining

Classes: This is used to locate textual data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they usually order. This information could be used to increase traffic by having daily specials.

Clusters: Text data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify consumer affinities or market segments.

Associations: Text can be mined to identify associations. The market basket analysis is an example of associative mining.

Sequential patterns: Text is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a knapsack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

II. TEXT MINING and DATA MINING

This section discusses the relationship between text mining and data mining. Although both the terms can be interchangeably used at many places but there are certain

issues at which they contrast each other. Data mining can be described as looking for patterns in data whereas text mining is about looking for patterns in text. However, the posing similarity between the two conceals real differences. Data mining can be characterized as the extraction of implicit, previously unknown, and potentially useful information from data. The information is implicit in the input data. It is hidden, unknown, and could hardly be extracted without recourse to automatic techniques of data mining. With text mining, however, the information to be extracted is clearly and explicitly stated in the text. It's not hidden at all, the only sense in which it is "previously unknown" is that human resource restrictions make it infeasible for people to read the text themselves. Text mining strives to bring it out of the text in a form that is suitable for consumption by computers directly, with no need for a human intermediary. Though there is a clear difference theoretically, from the computer's point of view the problems are quite similar. Text is just as dense as raw data when it comes to extracting information. Another requirement that is common to both data and text mining is that the information extracted should be potentially useful. In many data mining applications, potentially useful is given a different interpretation: the key for success is that the information extracted must be comprehensible in that it helps to explain the data. This is necessary whenever the result is intended for human consumption rather than a basis for automatic action. This criterion is less applicable to text mining because, unlike data mining, the input itself is comprehensible. Text mining with comprehensible output is synonymous to summarizing salient features from a large body of text, which is a subfield in its own right i.e. text summarization.

III. WORKING of TEXT MINING

Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining. These various stages of a text-mining process can be combined into a single workflow.

A. *Information Retrieval (IR) Systems*

It identify the documents in a collection which match against a user's query. The most well-known IR systems are search engines such as Google, which is used to identify those documents on the World Wide Web that are relevant to a set of given words. Information Retrieval systems are repeatedly used in libraries, where the documents are typically not the books themselves but digital records containing information about the books. This is however changing with the arrival of digital libraries, where the documents being retrieved are digital versions of books and journals. Information Retrieval systems allow us to minimise the set of documents that are relevant to a particular problem. As text mining involves applying very computationally concentrated algorithms to large document collections, IR can speed up the analysis considerably by reducing the number of documents for analysis.

B. *Natural Language Processing (NLP)*

It is one of the most difficult problems in the field of artificial intelligence. It is the analysis of human language so that computers can understand natural languages just like humans do. Although this goal is still miles away, NLP can perform some types of analysis with a high degree of success. For example: (1) Part-of-speech tagging which classifies words into categories such as noun, verb or adjective, (2) Word sense disambiguation which identifies the meaning of a word, given its usage, from the multiple meanings that the word may have, (3) Parsing that performs a grammatical analysis of a sentence. The role of NLP in text mining is to provide the systems in the information extraction phase with linguistic data that they need to perform their task.

C. *Information Extraction (IE)*

It is the process of automatically obtaining structured data from an unstructured natural language document. It involves defining the general form of the information that we are interested in as one or more templates, which are then used to monitor the extraction process. IE systems rely heavily on the data generated by NLP systems. Tasks that IE systems can perform include: (1) Term analysis, which identifies the terms in a document, where a term may consist of one or more words, (2) Named-entity recognition, which identifies the names in a document, such as the names of people or organisations, (3) Fact extraction, which identifies and extracts complex facts from documents. The data generated during IE are normally stored in a database ready for analysis in the final stage, data mining.

D. *Data Mining (DM)*

It is the process of identifying patterns in large sets of data. The aim is to uncover useful knowledge which was previously unknown. When used in text mining, DM is applied to the facts generated by the information extraction phase. We put the results of our DM process into another database that can be queried via a suitable graphical interface by the end-user. The data generated by such queries can also be represented visually.

IV. TEXT MINING TECHNIQUES

A. Classification

This is the most widely used text mining technique. Most decision making models are usually based upon classification methods. These techniques are also called classifiers that enable the categorisation of data into pre-defined classes. The

use of classification algorithms involves a training set which consists of pre-classified examples and are used to determine a set of parameters required for proper discrimination between the classes. The algorithm then encodes these parameters into a model called a classifier. There are many algorithms that can be used for classification, such as decision trees, neural networks, logistic regression, etc. Using this mining technique, the text mining tool learns from the data that how to partition or classify certain objects. It can be an object, an action, or any other information, that can be formalised. As a result, data mining software formulates classification rules.

Example - customer database
Question - Does the customer belong to loyal group?
Typical rule formulated is -
if PROFIT > Rs10000 and PURCHASED = monthly and INCIDENTS = 0 then CUSTOMER_TYPE = LOYAL

B. Clustering (Segmentation)

This mining technique is used to explore groupings within data or entities. This approach is mainly used for segmentation. Clustering method allows entities to be partitioned into individual groups which also called segments. The main difference between classification and clustering is that clustering is structuring data without knowing anything about classes, while classification method assigns new knowledge to the classes that are known prior. Clustering algorithms can be classified to: The type of data input to the algorithm, the clustering criterion defining the similarity between data points, the theory and fundamental concepts on which cluster analysis techniques are based. The algorithms can be classified into the following types,

1) Partitional Clustering:

Partitional clustering decomposes the data set into a set of disjoint clusters. More specifically, they determine an integer number of partitions that optimize as certain criterion function. The criterion function may accentuate the local or global structure of the data and its optimization is an iterative procedure. Given a database of n data tuples, a partitioning method constructs 'k' of the data, where each partition represents a cluster and k<= n. Thus it classifies the data into k groups, which together satisfy the following requirements: Each object must belong exactly one object and each object must belong to one group. The applications adopt one of two popular heuristic methods, k-Means algorithm, where each cluster is represented by the mean value of the objects in the cluster and k-Medoid algorithm, where each cluster is represented by one of the objects located near the centre of the cluster. These clustering methods work well for finding spherical shaped clusters in small to medium sized databases.

2) Hierarchical Clustering:

The hierarchical clustering proceeds successively by either merging smaller clusters into larger ones or by splitting larger clusters into smaller ones. The result of this algorithm is a tree of clusters which is called dendrogram that shows how the different clusters are related to each other. By cutting the dendrogram at desired level, a clustering of the data items into disjoint groups is obtained. Based on how the hierarchical decomposition is formed, a hierarchical method can be either agglomerative or divisive. The agglomerative approach also called the bottom-up-approach, starts with each object forming a separate group. It successively merges the objects or groups close to one another, until all the groups are merged into one. The divisive approach also called the top-down approach, starts with all objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition meets.

3) Density Based Clustering:

This clustering is used to group neighbouring objects of a data set into clusters based on density conditions. DBSCAN is a typical density-based method that grows cluster according to density threshold. OPTICS is another density-based method that computes an augmented clustering ordering for automatic and interactive cluster analysis.

4) Grid Based Clustering:

This type of algorithm is primarily proposed for spatial data mining. Their main characteristics is that they quantize the space into finite number of cells and then they do all operations on the quantize space. STING is a typical example of Grid - based method.

C. Association

Association rules are basic types of patterns or regularities that are found in transactional-type data. This data mining technique has its origins in traditional retail marketing where it can determine similarities between items that occur within a particular shopping mall for example, what items typically co-occur as contents of a shopping basket. Hence, a substitute name for this type of analysis is "market-basket analysis". For two sets of characteristics C1 and C2, an association rule is usually denoted as to convey that the presence of the characteristic C1 in a transaction often implies the presence of characteristic C2. With the help of association methods mining software creates rules that associate one attribute of a relation to another. Discovering these rules is very efficient on set oriented approaches. For example consider customer database in a supermarket
75% of customers who purchase Product1 also purchase Product2 where 75 is the confidence factor of the rule, shows the association between Product1 and Product2.

D. Sequence/Temporal

Sequential patterns involve mining frequently occurring patterns of activity over a period of time. In many situations, where not only the coexistence of items within a transaction be important but also the order in which those items appear across ordered transactions and the amount of time between transactions is considered, this technique proves to be very useful. Thus sequential pattern detection methods are similar to association rules except that they look for patterns across time. This could be a pattern that represents a sequence of tax filings over time or a sequence of purchases over time, etc. Sequence rules differ from other mining methods with the temporal factor.

V. APPLICATIONS

This technology is now broadly applied for a wide variety of government, research, and business needs. Here we briefly discuss some of the applications of text mining methods in different areas as patent analysis, text classification in news agencies, sentiment analysis, spam filtering and so on. Each of the applications has specific characteristics that need to be considered while selecting appropriate text mining methods.

A. Patent Analysis

In recent years the analysis of patents developed to a large application area. Many supervised and unsupervised techniques are applied to analyse patent documents and to support companies. The main challenges in patent analysis consists of the length of the documents, which are larger than documents usually used in text classification. Commonly every document consist of 5000 words in average. Thus these methods have been used by various commercial products in many ways but are still of interest for research, because there is still a need for improved performance. Many well-known companies offer products to support the analysis of patent text documents.

B. Text Classification for News Agencies

In publishing houses a large number of news stories arrive every day. These stories should be tagged with categories and the names of important persons, organizations and places so that required knowledge could be drawn easily. To automate this process a commercial text classification system was selected to support the annotation of news articles.

C. Anti-Spam Filtering of Emails

The growth of unsolicited e-mail or spam over the last few years has been increased rapidly. One solution is offered by anti-spam filters. Most commercially available filters use black-lists and hand-crafted rules. On the other hand, the success of machine learning methods in text classification offers the possibility to arrive at anti-spam filters that quickly may be adapted to new types of spam.

D. Sentiment analysis

Sentiment analysis may be used for the analysis of movie reviews for estimating how favorable a review is for a movie. Such an analysis may need a labeled data set or labeling of the affectivity of words. Text has been used to detect emotions in the related area of affective computing. Text based approaches to affective computing have been used on multiple corpora 10.

E. Academic applications

The topic of text mining is important to publishers who hold large databases of information that needs indexing for retrieval. This is especially helpful in scientific disciplines, in which highly specific information is often contained within written text. Therefore, initiatives have been taken that would provide semantic cues to machines to answer specific queries contained within text without removing publisher barriers to public access.

I. References

[1] Subbaiah , S. "Extracting Knowledge using Probabilistic Classifier for Text Mining " Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22.

[2] Madhusudhan, C.H., Rao K. Mrithyunjaya "Proposed Architecture for Automatic Conversion of Unstructured Text Data into Structured Text Data on the Web" IJCSNS International Journal of Computer Science and Network Security, VOL.13 No.12, December 2013

[3] http://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf

[4] Selvakumar, A. "An Adaptive Partitional Clustering Method for Categorical Attribute using K-Medoid" IJCSMC, Vol. 2, Issue. 4, April 2013, pg.197 – 204

[5] Berry, Michael W. "survey of text mining clustering classification and retrieval" http://books.google.co.in/books?hl=en&lr=&id=WNxbDFbmO-8C&oi=fnd&pg=PR11&dq=classification+of+words+using+text+mining&ots=YTZExbgD5s&sig=FH_PX22I2QhKh2Ey3rS4a5pusXE#v=onepage&q=classification%20of%20words%20using%20text%20mining&f=false

[6] http://en.wikipedia.org/wiki/Text_mining

[7]http://sitecore.jisc.ac.uk/media/documents/publications/bptextminingv2.pdf

[8] www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf

