IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 5, Ver. IV (Sep.- Oct. 2017), PP 16-19
www.iosrjournals.org

A Survey on Techniques And Its Applications of Text Mining

Swathi Agarwal[1], G.L.Anand Babu[2], G.Sekhar Reddy[3]
[1,2,3]*Department Of Information Technology, Anurag Group Of Institutions, Hyderabad, Telangana, India.*

***Abstract**: Text mining is a procedure that utilizes an arrangement of algorithms for changing over unstructured content into organized data items and the quantitative techniques used to break down these data items. The principal target of Text mining is to empower clients to separate information from text based resources and deals with the operations like recovery, extraction, rundown, order (directed) and grouping (unsupervised). Keeping in mind the end goal to locate a productive and compelling system for text classification, different methods of text categorization is recently developed. Some of them are regulated and some of them unsupervised way of report course of action. In this paper, focus is text mining process, diverse technique for text classification, group examination for content reports and its applications.*

***Keywords**: Clustering, Information Extraction, Information Retrieval, Natural Language Processing, Natural Language Text, Query Processing, Text mining.*

Date of Submission: 20-09-2017 Date of acceptance: 10-10-2017

I. Introduction

Average text mining assignments incorporate text categorization, text clustering, and idea/substance extraction, creation of granular scientific categorizations, assessment investigation, record outline, and element connection displaying (i.e., learning relations between named elements). Text mining is unique in relation to what are generally comfortable with in web mining. At the point when client seeks something in web that is as of now known and which is composed by another person. The primary issue in web mining is obtaining all materials, which are not important to our pursuit and in addition it won't show obscure data yet in text mining the fundamental objective is to find the obscure data something that nobody knows [1].

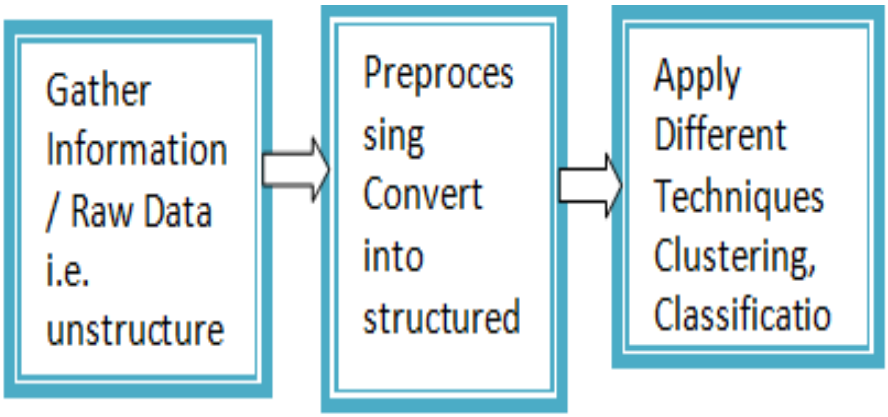

Figure: Basic Process of Text Mining

II. Processing Of Text Mining

There are five basic text mining steps as under:

Text mining steps:

a) Collecting information from unstructured data.
b) Convert this information received into structured data
c) Identify the pattern from structured data
d) Analyze the pattern
e) Extract the valuable information and store in the database.

Text mining involves a series of activities to be performed in order to efficiently mine the information. These activities are:

Text Pre-processing:
(a). Text Cleanup:
Text Cleanup implies evacuating of any pointless or undesirable data, for example, expel advertisements from site pages, standardize content changed over from double arrangements, manage tables, figures and recipes.

(b). Tokenization:
Tokenizing is essentially accomplished by part the content on blank areas and at accentuation denotes that don't have a place with shortenings distinguished in the first step.
(c). Part of Speech Tagging:
Grammatical form (POS) labeling implies word class task to every token. Its information is given by the tokenized content. Taggers need to adapt to obscure words (OOV issue) and questionable word-label mappings. Manage based methodologies like ENGTWOL [8] work on a) lexicons containing word frames together with the related POS names and morphological and syntactic highlights and b) setting touchy tenets to pick the suitable names amid application.

Text Transformation (Attribute Generation):
A content archive is spoken to by the words (highlights) it contains and their events. Two principle methodologies of record portrayal are a) Bag of words b) Vector Space.
Feature Selection (Attribute Selection):
Highlight choice otherwise called variable determination, is the way toward choosing a subset of essential highlights for use in display creation. The primary presumption when utilizing a component determination system is that the information contains numerous repetitive or unimportant highlights.
Data Mining:
Now the Text mining process converges with the conventional Data Mining process. Great Data Mining procedures are utilized as a part of the organized database that came about because of the past stages.
Evaluate:
Assess the outcome, after assessment the outcome can be disposed of or the created result can be utilized as a contribution for the next set of sequence.

III. Text Mining Technologies

The main purpose of text mining techniques is to structure the text documents. The following are the important text mining techniques.
1. Information Extraction
2. Information retrieval
3. Natural Language processing
4. Query processing
5. Clustering

1. Information Extraction (IE)

Information Extraction (Winter School) is a procedure of consequently separating organized data from unstructured or semi-organized natural language text. Pattern matching is the last yield of the extraction procedure. It is the sort of database acquired by searching for predefined groupings in text. This includes characterizing the general type of the data that we are keen on as at least one formats, which are then used to control the extraction procedure.

2. Information retrieval

Information Retrieval (IR) frameworks distinguish the records in an accumulation which coordinate a client's question. The most critical use of data recovery is web crawler in World Wide Web, which recognize those archives on the WWW that are essential to an arrangement of given words. The way toward discovering data as indicated by the client's demand is data recovery. Ordinarily, it alludes to the programmed recovery of archives. Information retrieval deals with crawling, indexing document and retrieving document [2]. It is utilized to recover accumulation of noteworthy pages from the arrangement of pages in WWW. Database framework contracts with query based organized information. Information Retrieval manages inquiry in view of extensive measure of content reports.

3. Natural Language processing

The general purpose of NLP is to achieve a better understanding of natural language by use of computers. The range of the assigned techniques reaches from the simple manipulation of strings to the automatic processing of natural language inquiries.

4. Query processing

Once a modified record is made for an archive accumulation, a recovery framework can answer a keyword query rapidly by looking into which reports contain the query keywords. In particular, we will keep up a score gatherer for each report and refresh these collectors as we experience each query term. One noteworthy impediment of many existing retrieval techniques is that they depend on correct keyword matching.

5. Clustering

The procedure in which objects of intelligently comparative properties are physically set together in one class of articles and a solitary access to the circle makes the whole class accessible is clustering. This strategy is utilized to aggregate comparable archives, yet it contrasts from arrangement that records are bunched on the fly rather than using predefined themes. Another advantage of grouping is that archives can show up in numerous subtopics, in this way guaranteeing a valuable report won't be overlooked from list items [2]. A fundamental bunching calculation makes a vector of subjects for each archive and measures the weights of how well the report fits into each group. Clustering technology can be helpful in the association of administration data frameworks, which may contain a huge number of reports.

IV. Text Mining Applications

Text mining is a generally new region of software engineering, and its utilization has developed as the unstructured information accessible keeps on expanding exponentially in both pertinence and amount [3]. Text mining can be utilized to make the extensive amounts of unstructured information open and helpful, accordingly creating esteem, as well as conveying ROI from unstructured information administration as we've seen with utilizations of content digging for Risk Management Software and Cybercrime applications.

1 – Risk management

Deficient hazard investigation is regularly a main source of disappointment. This is particularly valid in the monetary business where selection of Risk Management Software in view of text mining innovation can drastically expand the capacity to moderate hazard, empowering complete administration of thousands of sources and petabytes of content archives, and giving the capacity to connect together data and have the capacity to get to the correct data at the perfect time.

2 – Knowledge management

Not having the capacity to discover imperative data rapidly is dependably a test while overseeing extensive volumes of content records—simply ask anybody in the social insurance industry. Here, information administration programming in light of text mining offer an unmistakable and dependable answer for the "data overabundance" issue [4].

3 – Customer care service

Text mining, as well as natural language processing is frequent applications for customer care. Today, text examination programming is habitually received to enhance client encounter utilizing distinctive wellsprings of important data, for example, studies, inconvenience tickets, and client call notes to enhance the quality, viability and speed in settling issues.

V. Conclusion

In this proposed framework, the attention to text mining strategies have been acquainted and displayed. Attributable to its uniqueness, there are numerous imminent research territories in the field of Text Mining, which involves the disclosure of better intermediary forms for demonstrating the yields of information extraction or recovery. The fixation has been determined on essential techniques for coordinating text mining. Text mining strategies are utilized to inspect the fortifying and related data brilliantly and capably from huge amount of indistinct information. This paper offers a fleeting synopsis of text mining methods that help to recuperate the text mining process. This paper upgrades the procedure and employments of text mining, which can be connected in huge number zones, for example, web mining, therapeutic, continue filtration, and so forth.

References

[1]. Shilpa Dang, Peerzada Hamid Ahmad, "Text Mining: Techniques and its Application", International Journal of Engineering & Technology Innovations, ISSN (Online): 2348-0866, Volume 1, Issue 4, pp. 22-25, 2014.
[2]. Swathi Agarwal, G. L. Anand Babu, Dr. K. S. Reddy, "Classification Techniques in Data Mining-Case Study", International Organization of Scientific Research, Volume 18, Issue 6, 2016.
[3]. Shah Neha K, "Introduction of Text mines and an Analysis of Text mining Techniques", PARIPEX, ISSN: 2250-1991, Volume 2, Issue-2, 2013.
[4]. T. Nasukawa, T. Nagano, "Text analysis and knowledge mining system", IBM Systems Journal, 2001, Volume: 40, Issue: 4, Pages: 967 – 984.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Swathi Agarwal. "A Survey on Techniques And Its Applications of Text Mining." IOSR Journal of Computer Engineering (IOSR-JCE) , vol. 19, no. 5, 2017, pp. 16–19.