RESEARCHERID
THOMSON REUTERS
[Girdher, 4(6): June 2017] ISSN 2348 – 8034
DOI- 10.5281/zenodo.817359 Impact Factor- 4.022
GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

TEXT MINING TECHNIQUES-A REVIEW

Heena Girdher*[1] & Poonam Gaur[2]
*[1,&2]Assistant Professor, Department of Computer Applications,Chandigarh Group of colleges, Landran (Mohali), India

ABSTRACT

Text mining is a technology that is used to extract meaningful information from unstructured or semi structured text. The amount of data is increasing at tremendous speed. So there is a need to extract meaningful information from huge amount of data. Text mining techniques are used for this purpose. This paper focuses on text mining process, various techniques of text mining. In addition to this we have also discussed a comparison between text mining techniques on the basis of Goal, Algorithms and Tools.

Keywords*: Classification, Clustering, Information Extraction, Information Retrieval, Summarization, Text Mining*.

I. INTRODUCTION

Recent years have witnessed the rapid growth of web, which is the main source of the data. The amount of data is increasing at tremendous speed. Data may be structured, unstructured or semi structured. Structured data concerns all data which can be stored in SQL in table with rows and columns but structured data represent only 5 to 10% of the data. Semi structured data is information that does not reside in relational database. Xml, Json, NOSQL databases are considered as semi structured. Unstructured data represents 80% of the data [1]. It includes emails, word documents, power point presentations, web pages and instant messages.

We are drowning in data but starving for information. There is a need to extract meaningful information from huge amount of data. Data mining deals with extracting meaningful information from structured text. Text mining is the technique of data mining which deals with extracting information from semi structured or unstructured text.

Text mining [2] is the process of deriving high quality information from semi structured or unstructured data. High quality information is acquired by finding patterns and trends through means such as statistical pattern learning. Text mining is also named as text data mining or text analytics.

Text mining is not same as keyword search. Traditional keyword search gets all the documents that contain the keywords you have specified but you still have to read all those documents to find the relevant information. Text mining software reads and analyzes the document on your behalf.

[Girdher, 4(6): June 2017] ISSN 2348 – 8034
DOI- 10.5281/zenodo.817359 Impact Factor- 4.022

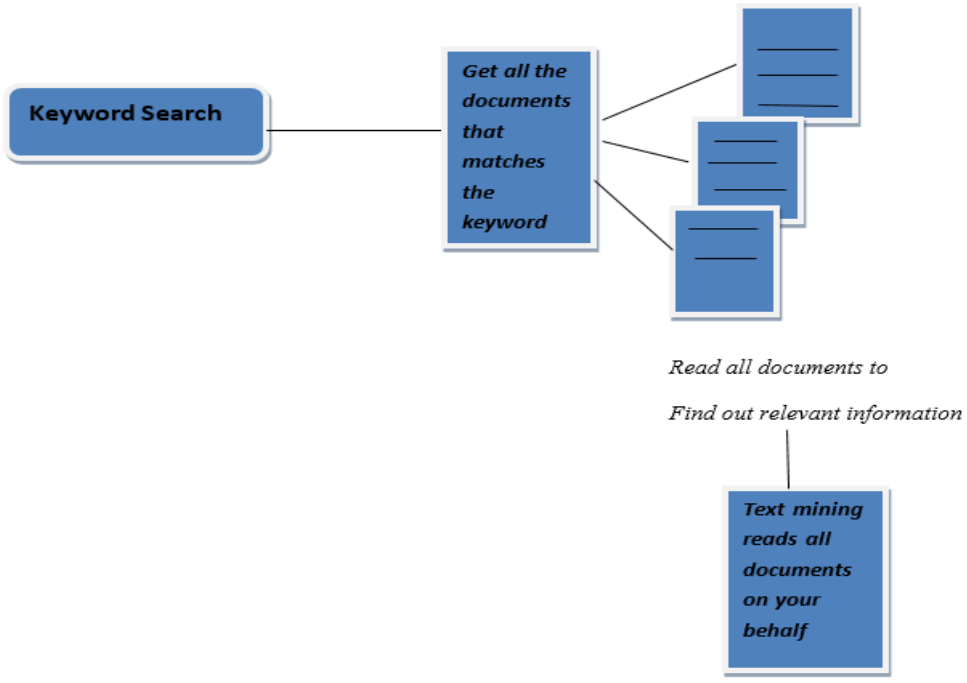

Figure 1.1: Comparison between Text Mining and Keyword Search

II. TEXT MINING PROCESS

Text mining act as a step in the process of knowledge discovery.

i. **Text Document Gathering: -** Text documents are collected from many different resources. Text documents may be in the form of pdf, word document and web page etc.
ii. **Text Preprocessing: -** Text document may contain unwanted and noisy data. So there is a need to preprocess the data. Text preprocessing involves following steps:-
iii. **Text Cleanup: -** Text document is cleaned to remove unwanted information like remove ads from web pages.
iv. **Tokenization: -** Text document is considered as a string. The whole document is divided into tokens separated by delimiter.
v. **Removal of Stop words: -** Stop words are meaningless words which doesn't effects the meaning of text like a, an, the, but, of etc. Stop words areremoved from the text to reduce the size of text.
vi. **Stemming: -** Stemming is the processof reducing a word to its root word. For example jumps and jumped may be reduced to jump. The most common algorithm for stemming is Porter's algorithm.
vii. **Text Transformation: -** Text document is transformed into forms that are appropriate for mining. Text document istransformed into vector space model or bag of wordsapproach for further effective analysis.
viii. **Feature Selection: -** Text document may contain relevant and irrelevant features. In feature selection procedure, irrelevant features are removed from the text to obtain a reduced representationof the text to reduce computation process.
ix. **Text mining techniques: -** Text mining techniques combine with the data mining techniques are applied to the text to obtain patterns.
x. **Evaluation: -** The obtained patterns are evaluated according to interestingness measures [3].

[Girdher, 4(6): June 2017] ISSN 2348 – 8034
DOI- 10.5281/zenodo.817359 Impact Factor- 4.022

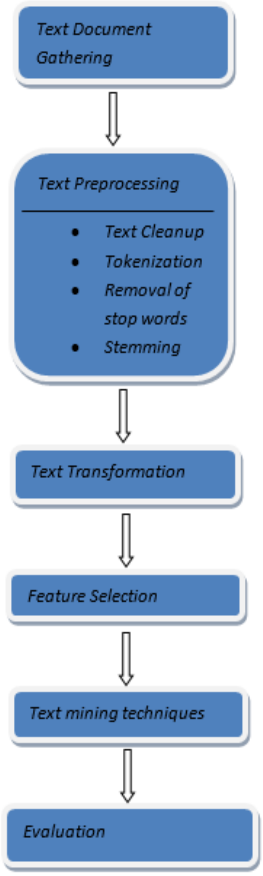

Figure 1.2: Text Mining Process

III. TEXT MINING TECHNIQUES

There are various text mining techniques discussed below:-

i. Information Extraction:

Information extraction [4], [10] is the process of extracting structured information from unstructured text for analysis. The structured information involves extraction of entities like name of person, location and organization, relationship between entities like " is employee of" relationship between a person and an organization, "is acquired by" relationship between a pair of companies, opinion of entities which may be positive or negative etc.

ISSN 2348 – 8034
Impact Factor- 4.022

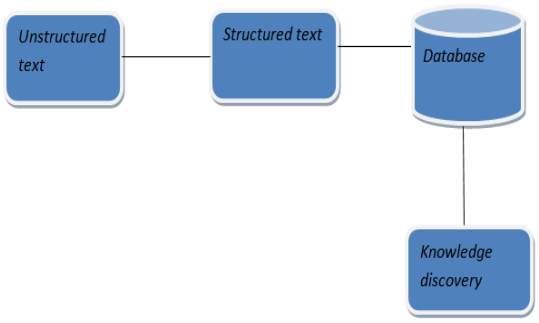

Fig 1.3: Information Extraction

Applications

- News tracking: - It involves automatically tacking of event types from news articles.
- Sentiment analysis:-It involves extraction of opinions of users from social networking sites for better decision making.
- Information extraction from emails
- Person profile extraction
- Information extraction in Digital libraries
- Table extraction using conditional Random fields

ii. Information retrieval: -

Information retrieval [5], [8] is the process of extracting the relevant information from the internet. The user types a query on the searchengine and the document relevant to the query are extracted. Document retrieval is considered as an extension of Information retrieval where the documents that are returned are processed to find the relevant information.

Applications

- Digital Library
- Recommender system
- Search engines
- Media Search

iii. Summarization:

Summarization [7] is the process of creating reduced representation of large amount of data that list the main point of the original document. The amount of data is increasing at growing rate. So it is difficult to create the reduced representation of data manually. Therefore there is a need for automatic summarizer software which summarize the document by itself. In document summarization, a summary of the document is created by finding the most relevant sentences whereas in image summarization most relevant part of the image is find out.

Applications:-

- News feed
- Report generation
- Mail clients
- Sentence compression
- Entity timelines
- Storylines of events
- Summarization of user generated content

iv. Clustering

Clustering [6], [10] is the process of grouping data items of similar type into one cluster and dissimilar type into another cluster. It is an unsupervised learning process.

To understand the concept of clustering, let us take an example of a supermarket. In a supermarket, items are grouped according to similarity. Items of same type are placed in same section. Suppose a user is interested in Crockery items, she will directly go to the kitchen section instead of searching it in the whole supermarket. So clustering is useful in searching items in less time. Clustering can also be used for outlier detection. The objects that do not fit in any cluster are called outliers. Clustering can also be used for credit card fraud detection. Document clustering organize the large amount of documents into clusters based on some similarity It can be used to browse a collection of documents or to organize the result returned by search engine in response to a user's query.

Applications

- Data mining
- Pattern recognition
- Image analysis
- Bio Informatics
- Taxonomy generation
- Topic extraction

v. Classification

Classification is the process of finding the class label of a new tuple. Suppose a manager wants to determine whether the customer will buy the computer or not. The classes for buys computer will be yes or no. In classification, a classification model or classifier is constructed from given training data and the class label of the test data is determined.

Text classification [9] is the process of assigning a category to a new text document. Text classifier is used to categorize the text document into predefined class.

Applications

- Business
- Medicine
- Law
- Society

Table 1.1: Comparison of various text mining techniques

	Goal	Algorithms	Tools
Information retrieval	Find document relevant to an information need from a large document set	1. Retrieval algorithm 2. Filtering algorithm 3. Indexing algorithm	Intelligent Miner, Text Analyst
Classification	To build a set of models that can predict the class of different objects	1. KNN algorithm 2. Naïve Bayes 3. Concept Vector based algorithm 4. Decision tree induction	Alceste, Monkeylearn
Information extraction	Extracting structured information from unstructured or semi structured documents	1. RAPIER 2. BWI 3. MBL 4. TBL	Text Finder, Clear Forest Text

ISSN 2348 – 8034
Impact Factor- 4.022

		5. SVM	
Clustering	Group the items into k clusters such that all items in same cluster are similar to each other as possible and items	1. K-means 2. Hierarchical 3. K-medoids 4. DBSCAN	Clustify, Carrot, Rapid Miner
Summarization	To create a summary with the major points of the original document.	1. Summarizer 2. Text Rank 3. Lex Rank	Tropic tracking tool, Sentence Ext tool

IV. CONCLUSION

In this review we have described the text mining process. In addition to this various techniques of text mining such as information retrieval, information extraction, summarization, classification, clustering have been introduced and presented. Additionally, we have discussed the comparison of text mining techniques on the basis of Goal, Algorithms and Tools

REFERENCES

1. *Vishal Gupta and Gurupreet Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Volume 1, No. 1, 2009.*
2. *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.*
3. *R. Balamurugan, Dr. S. Pushpa, "A Review on various Text Mining Techniques and Algorithms", International Journal of Advance Research in Science and Engineering, Volume-4, Issue 11, 2015*
4. *Peerzada Hamid Ahmad, Shilpa Dang, "A Comparative Study on Text Mining Techniques", International Journal of Science and Research, ISSN: 2319-7064, Volume 3, Issue 12, pp. 2222-2226, 2014*
5. *Qing Cao, Wenjing Duan, Qiwei Gan, "Exploring determinant s of voting for the "helpfulness" of online user reviews: A text mining approach', 0167-9236/$ – see front matter © 2010 Elsevier B.V. All rights reserved. doi:10.1016/j.dss.2010.11.009*
6. *Lokesh Kumar and Parul Kalra Bhatia,"TextMining:Concept Process, Applications," Journal of Global Research in Computer Science Volume 4, No. 3, March 2013 .*
7. *Dr shilpa Dang, Peerzada Hamid Ahmad, "A Review of Text Mining Techniques Associated with Various Application Areas", International Journal of Science and Research, ISSN (Online): 2319-7064, Volume 4, Issue 2, 2015*
8. *Vidhya. K. A and G. Aghila, "Text Mining Process, Techniques and Tools: an Overview", International Journal of Information Technology and Knowledge Management, Volume 2, No. 2, pp. 613-622, 2010.*
9. *Chauhan Shrihari R, Amish Desai, "A Review on Knowledge discovery using Text classification techniques in Text Mining", International Journal of Computer Applications (0975-8887) Volume-111-No 6,2015*
10. *Varsha C. Pande and A.S. Khandelwal "A Survey of Different Text Mining Techniques", IBMRD's Journal of Management & Research, ISSN: 2348-5922, Volume 3, No. 1, pp. 125-133, 2014*