Volume: 08, Issue: 05 Pages: 82-85 (2017) Special Issue

A Study of Various Text Mining Techniques

D. Jasmine Guna Sundari
Assistant professor
Department of Computer Science, Government Arts College for women, Ramanathapuram
Email: durkarthi@gmail.com
D. Sundar
Assistant professor
Department of Computer Science, Government Arts College, Thiruvadanai
Email: sundarums@gmail.com

ABSTRACT

Text mining is an exciting research area that tries to discover useful information can be derived from this unstructured data by using techniques from machine learning, natural language processing (NLP), data mining, information retrieval (IR), and knowledge management. Text mining involves the pre-processing techniques to harvest data and initial understanding of the patterns that exist in the data. The techniques such as Information Retrieval, Information Extraction, Categorization, Clustering and Summarization that are used to analyse these intermediate representations such as distribution analysis, association rules and visualisation of the results.

Keywords **- Information Retrieval, Extraction, Text categorization, Clustering, Summarization, Visualization.**

1. INTRODUCTION

Massive amount of new information being created, above 80-90% of all data is held in various unstructured formats. The text mining extracts the useful information from data sources through the explorations and identifications of interesting patterns. In this case of text mining, the data sources are document collections and patterns are not found among formalised database records but in the unstructured textual data in the documents in these collections. Text mining is a multidisciplinary field, concerning retrieval of information, analysis of text, extraction of information, categorization, clustering, visualization, mining of data, and machine learning. Text mining can work with unstructured or semi-structured data sets such as emails, full text documents and HTML documents etc.

Text Mining = Statistical NLP + Data Mining

Statistical NLP A set of algorithms for converting unstructured text into structured data objects.

Data Mining The quantitative methods that analyze these data objects to discover knowledge.

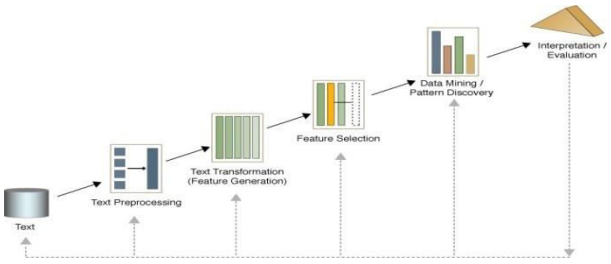
Text mining steps

1) Extract information from unstructured data.
2) Extracted information converted into structured data.
3) Pattern identified from structured data.
4) Analyze the pattern.
5) Extract the valuable information.
6) Store in the database.

2. TEXT MINING PRE-PROCESSING TECHNIQUES

There are two ways of categorizing the structuring techniques of document are according to their task, algorithms and formal frameworks that they use. Task oriented pre processing approaches envision the process of creating a structured document representation in terms of tasks and subtasks and usually involve some sort of preparatory goal or problem that needs to be solved such as extracting titles and authors from a PDF. In pre processing approaches are rely on techniques such that classification schemes, probabilistic models, and rule-based systems approaches for analysing complex phenomena that can be also applied to natural language texts.

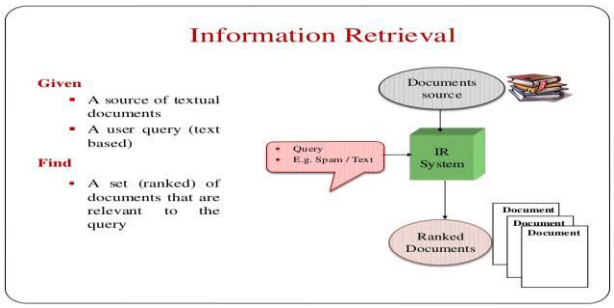
- Text Preprocessing
  - Syntactic/Semantic text analysis
- Features Generation
  - Bag of words
- Features Selection
  - Simple counting
  - Statistics
- Text/Data Mining
  - Classification (Supervised) / Clustering (Unsupervised)
- Analyzing results



3. INFORMATION RETRIEVAL

Information retrieval system is a network of algorithms, which facilitate the search of relevant data / documents as per the user requirement. It not only provides the relevant information to the user but also tracks the utility of the displayed data as per user behaviour, i.e. is the user

finding the results useful or not? The most well known information retrieval (IR) systems are Google search engines which recognize those documents on the World Wide Web that are associated to a set of given words. It is measured as an extension to document retrieval where the documents that are returned are processed to extract the useful information crucial for the user. Thus document retrieval is followed by a text summarization stage that focuses on the query posed by the user, or an information extraction stage. IR in the broader sense deals with the whole range of information processing, from information retrieval to knowledge retrieval. It gained increased attention with grow of the World Wide Web and the search engines.



4. INFORMATION EXTRACTION

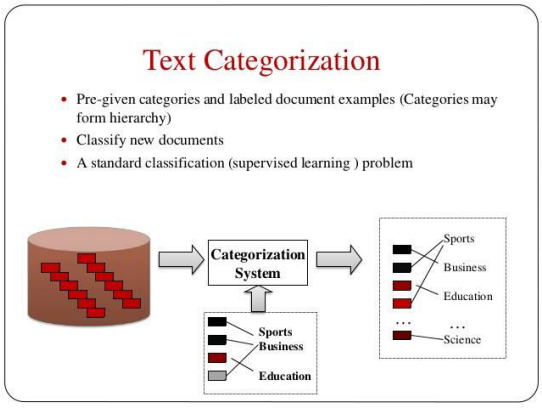
The main goal of information extraction (IE) methods is the extraction of useful information from text. It identifies the extraction of events, entities and relationships from semi-structured or unstructured text. Information extraction software identifies key phrases and relationships within text. IE is concerned with extraction of semantic information from the text. The extraction algorithm which attempts to improve recall by using the mined rules is summarized in figure.

Input *RB is the set of prediction rules.*
D is the set of documents.
Output *F is the set of slot fillers extracted.*
Function *InformationExtraction (RB, D)*
F: = Ø.
For each example D*D do*
Extract fillers from D *using extraction rules and add them to F*
For each rule R in the prediction rule base RB do
If R fires on the current extracted fillers
If the predicted fillers is a substring of D
Extract the predicted filler and add it to F
Return F.

The Natural language texts have information, which is not suitable for computers for analysis purpose. Whereas computers uses large amount of text and extract useful information from passages, phrases or single words. So Information Extraction can be considered as restricted form of natural language understanding and here we know about the semantic information, we are seeking for. The task of information Extraction is to extract parts of text and assign specific attribute to it.

5. TEXT CATEGORIZATION

The goal of text categorization is to classify a set of documents into a fixed number of predefined categories; each document may belong to more than one class. The categorization task is to classify a given data instance into a pre-specified set of categories. Text categorization is a kind of "supervised" learning where the categories are known in advance and firm in progress for each training document. It involves identifying the main themes of the document into a pre-defined set of topics. When categorizing a document, a computer program will often treat the document as a "Bag of Words". Categorization is the assignment of normal language documents to predefined set of topics according to their content. It is a collection of text documents, the process of finding the accurate topic or topics for each document. Nowadays automated text categorization is applied in a variety of contexts from the classical automatic or semiautomatic indexing of texts to personalized commercials delivery, spam filtering, and categorization of Web page under hierarchical catalogues, automatic metadata generation. Categorization tools have a method for ranking the document in order of which documents have the most content on a particular topic.



6. CLUSTERING

Clustering is one of the most interesting and important topics in text mining which is used to group similar documents. Its aim is to find intrinsic structures in information, and arrange them into significant subgroups for further study and analysis. It is an unsupervised process through which objects are classified into groups called clusters. The problem is to group the given unlabeled collection into meaningful clusters without any prior information. Any labels associated with objects are obtained from the data. Another benefit of clustering is that documents can appear in multiple subtopics, thus ensuring that a useful document will not be omitted from search results. The related document to be retrieved once from one of the documents has been relevant to a query. A basic clustering algorithm creates vector topics for each document and measures the weights of how well the document fits into each cluster.

Clustering is useful in many application areas such as biology, data mining, pattern recognition, document retrieval, image segmentation, pattern classification, security, business intelligence and Web search.
The most commonly used clustering algorithms are the K-means (hard, flat, shuffling), the EM-based mixture resolving (soft, flat, probabilistic), and the HAC (hierarchical, agglomerative).

6.1 K-Means Algorithm

The K-means algorithm partitions a collection of vectors {x1, x2,..,xn} into the set of clusters {C1, C2, . . . Ck}. The algorithm needs k cluster seeds for initialization. They can be externally supplied or picked up randomly among the vectors.
The algorithm proceeds as follows
Initialization
K seeds, either given or selected randomly, form the core of k clusters. Every other vector is assigned to the cluster of the closest seed.
Iteration
The centroid Mi of the current cluster is computed
Each vector is reassigned to the cluster with the closest centroid.
Stopping condition
At convergence – when no more changes occur.
The K-means algorithm maximizes the clustering quality function Q, If the distance metric (inverse of the similarity function) behaves well with respect to the centroids computation, then each iteration of the algorithm increases the value of Q. A sufficient condition is that the centroid of a set of vectors be the vector that maximizes the sum of similarities to all the vectors in the set. This condition is true for all "natural" metrics. It follows that the K-means algorithm always converges to a local maximum.

The K-means algorithm is popular because of its simplicity and efficiency. The complexity of each iteration is O(kn) similarity comparisons, and the number of necessary iterations is usually quite small.

6.2 Hierarchical Agglomerative Clustering (HAC)

The HAC algorithm begins its work with each object in particular cluster and proceeds, according to some chosen criterion it is repeatedly merge pairs of clusters that are most similar. The HAC algorithm finishes when everything is merged into a single cluster. Binary tree of the clusters hierarchy is provided by history of merging.
The algorithm proceeds as follows:
Initialization
Each and every object is put into a separate cluster.
Iteration
Find the pair of most similar clusters and merge them.
Stopping condition
Repeat step 2 till single cluster is formed.
When everything is merged into single cluster different versions of the algorithm can be produced, then it is calculated the similarity between clusters. The complexity of this algorithm is O(n2s), where n is the number of objects and s the complexity of calculating similarity between clusters. Measuring the Quality of an algorithm needs human judgment, which introduces a high degree of subjectivity.
Given a set of categorized (manually classified) documents, it is possible to use this benchmark labeling for evaluation of clustering's. The most common measure is purity. Assume {L1, L2,..., Ln} are the manually labeled classes of documents, and {C1, C2, . . . , Cm} are the clusters returned by the clustering process.

7. SUMMARIZATION

Text summarization is helpful to reduce the length and detail of a document while retaining its main points and overall meaning. Text summarization is the process of automatically creating a compressed version of a given text that provides useful information for the user. Summarization tools may also search for heading and other markers of subtopics in order to identify the key points of a document. Microsoft word's AutoSummarize function is a simple example of text summarization.

An automatic summarization process can be divided into three steps

1) In the pre-processing step a structured representation of the original text is obtained.
2) In the processing step an algorithm must transform the text structure into a summary structure.
3) In the generation step the final summary is obtained from the summary structure.

Text summarization involves various methods that employ text categorization, such as neural networks, decision trees, semantic graphs, regression models, fuzzy logic and swarm intelligence. However, all of these methods have a common problem, that is, the quality of the development of classifiers is variable and highly dependent on the type of text being summarized.

8. VISUALIZATION

Visual text mining or information visualization puts large textual sources in a hierarchy or maps. The Information provided by graphical visualization is better, comprehensive and faster understandable than pure text based description so it is best for mining the large document collection. Information visualization is useful when a user needs to narrow down a broad range of documents and explore related topics. Most of the approaches of text mining are motivated by the methods which had been proposed in the area of visual data mining, information visualizations and explorative data mining.
This method can improve the discovery or extraction of relevant patterns or information for text mining and information retrieval systems. Information that allow a visual representation comprises aspects of result set, keyword relations or ontology are considered the aspects of the search process itself.

The goal of information visualization, the construction may be conducted into three steps

1) Data preparation

2) Data analysis and extraction
3) Visualization mapping

9. CONCLUSION

Text mining is the process of extracting valuable information from unstructured text. In this study, text mining techniques have been discussed. A comparison of different text mining techniques has been shown which can be further enhanced. 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. Now days there have been lot of work going on the document using text mining methods. The improvement for text mining is still an interesting, open issue and as in current world scenario time is the prime constraint of any application.

REFERENCES

[1]Shilpa Dangi, Peerzada Hamid Ahmad “Text Mining: Techniques and its Application”, IJETI International Journal of Engineering & Technology Innovations, Vol. 1 Issue 4, November 2014.

[2]Vishal Gupta, Gurpreet S. Lehal “A Survey of Text Mining Techniques and Applications, Journal of Emerging technologies in web intelligence, vol,1, no.1, August 2009.

[3]Pravin Shinde& Sharvari Govilkar “ A Systematic study of Text Mining Techniques”,International Journal on Natural Language Computing (IJNLC) Vol. 4, No.4, August 2015.

[4]Vidya K A, G Aghila, “Text Mining Process, Techniques and Tools: an Overview”, International Journal of Information Technology and Knowledge Management, July-December 2010, Volume 2,No 2, pp.613-622.

[5]R.Sagayam, S.Srinivasan, S.Roshini, “A Survey of Text Mining”Retrieval,Extraction and Indexing Techniques”. Internaltional Journal of Computational Engineering Research (ijceronline.com) Vol.2 Issue.5.

[6]Vishal Gupta and Guruprit Lehal, “A Survey of Text Mining Techniques and Applications”, Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, August 2009.

[7]Hearst, M. A. (1997) Text data mining: Issues, techniques, and the relationship to information access. Presentation notes for UW/MS workshop on data mining, July 1997.

[8]Rashmi Agrawal, Mridula Batra, "A Detailed Study on Text Mining Techniques", IJSCE, ISSN: 2231-2307, Vol. 2, Issue-6, January 2013.

[9]Falguni N. Patel, Neha R. Soni,"Text mining: A Brief survey", International Journal of Advanced Computer Research, ISSN (Online):2277-7970, Vol. 2, No. 4, Issue-6, Dec 2012.

[10]Mr. Rahul Patel,Mr. Gaurav Sharma,"A survey on text mining techniques", International Journal Of Engineering And Computer Science ISSN:2319-7242, Vol 3 Issue 5, May 2014, pp.5621-5625