International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Survey on Text Mining with Identification of Text Corpus

Swarupa Khapli

ME Student, Department of Computer Engineering, MITCOE, Pune, India

Abstract: ***Text mining is a term refers to text data mining. Text mining is the process of arranging unstructured data into structure data with statically analysis of texts. In the system, several records are the presence of large amounts of data in a distributed manner. Such data are stored in the form of files, record, documents. There is no proper way to manage. So the need of association of file is required. For showing the association of File gets from the text corpus using wordnet. Association of file, we can manage of file in a graphical way.***

Keywords: Text mining, text classification, wordnet, machine learning

1. Introduction

Text mining is an organization of text documents to classify the information in business insights from text-based content such as word documents, email, and chatting information gate from social media like as Facebook, Google talk, Twitter. Processing of unstructured information using natural language processing (NLP), statistical modeling and machine learning techniques can be challenging. It contains ambiguities caused by inconsistent syntax and semantics, including language specific to vertical industries and age groups. Typical text tasks involve the text categorizations, text clustering and documents summarizations.

For text identification is performed to get the relevant term from document [1]. For preprocessing of documents find the feature are examined the number of times words occur in a document. Multiple time word is converted in the form of indexing. Find some frequencies of the text from the documents. The document is like pdf, textetc. [7]. Using featured vector, find the similarity between multiple documents and showing the association between the documents.

Four Classifications of text, Different methods are used for preprocessing like to stop word removal, tf-IDF, streaming and for classification used feature selections and feature extraction. All this methodis described in Section 3 in details.

2. Literature Survey

The information obtained from various papers used for the text identification of co-relations. We are present the brief information about some of the paper.

A paper on topic identification in the text corpus by Chris Chilton proposed topic detection and tracking program. In that the mainly two problems specifying topic tracking for classifying the incoming new file and another one are topic detection for recognizing the new documents into an existing topic or belongs to the new topic [4]. For that purpose, data preparation, data cleaning, frequent itemset and clustering techniques are used. It is a combination of clustering on document mapping.

Another paper on sentence similarity based on semantic nets and corpus statistics proposed method derives text similarity from semantic and syntactic information combined in the compared text [3]. Semantic similarity between words techniques describes the hierarchical semantic knowledge base by lexical analysis using wordnet, gene ontology. Another technique is semantic similarity between sentences. These techniques form a dynamic semantic vector based on compare sentences. And another one paper on document classification based on word semantic hierarchies. In this paper document, classify based on the meaning of words of words and the relationships between concepts of group.

Another paper based on text classification using keyword extraction techniques proposed keywords extracted from documents using worn-out and tf-IDF method. Keyword extraction is using stopwords, elimination techniques. In that techniques remove the text that has less importance. Other techniques is stemming techniques are used to find out the root of the word. Another technique is the term frequency – inverse (Tf-IDF) [6] used as the weighting factor in information retrieval and text mining. Frequency has depended on the number of time words appears in the documents. Feature selection is the technique used after the preprocessing completed using tf-IDF. For classification used three machine techniques K-nearest neighbor, Naïve Bayes algorithm and decision tree [5].

3. Methods of Text Classifications

3.1 Keyword Extractions

A keyword is a term can be considered as summarized of a document in a short form. Keywords extractions are a technique performs the number of tasks related to text mining. Keywords extraction is used for to extract document retrieval, web page retrieval, documents clustering and summarizations. The important role of keyword extractions is to extract the word with relevance in the text. Keyword extraction is used for preprocessing of documents

3.2 Stop Word Eliminations

Stop word is a part of natural language and used for preprocessing of documents. They do not have meaning in a retrieval system. In Stop word elimination, least important word are removed from text corpus. Example of stop word is a, an, is, then, with etc. Stop words are removed the words from documents whose words are not considered as keywords in text mining.

3.3 Term Frequency- Inverse Document Frequency

Using Feature vector is to represent the documents. In that feature vector selection of one document as a set of term sequences, including term t and term weight w. Then the documents can be generates pairs of <t, w>, t1, t2, t3..tn represents features which describe the documents content. After that the creations of N-dimensional coordinate w1, w2, w3… wn represent the relevant coordinate with each other. So every document (d) mapped with respective feature vector v (d) = (t1, w1, t2, w2, t3, w3…. tn, wn)

The main role of data processing is to build up the data in the form of feature vector which deals with the data resources. For deciding the criteria, feature selection is to use the weight of the text content. The value of vector element represented as wi. Term Frequency (t, d) is the no of times words occurred in documents. Document frequency (t) is the number of documents in which the word occurs at once. The Inverse document frequency (t) can be calculated in the form of document frequency.

$$\text{IDF}(t) = \log\left[\frac{|D|}{DF(t)}\right]$$

|D| is the total number of documents. The find the value of wi of featureit for documents d is then calculated as the product of tf and IDF.
Wi = TF (ti,d) . IDF (ti)
Wi is the weight the word t, in the documents d.

3.4 Feature Selection

Feature selection is the process of selection of a subset or list of attributes or variables that are used to model describing data. The purpose of feature selections is to remove the dimensionality, irrelevant data, reducing the amount of data needed for learning, improving algorithms for predictive accuracy.

Dimension reduction techniques are the techniques based on Feature extraction and feature selections. Feature selection is the algorithms for select the subset of the most relevant features from original feature space. Feature Extraction is the algorithms to transform the original feature space to smaller feature space to reduce the dimensions. For effective dimension of data sets the text domain reduction.

The evaluation Functions for word frequency is

Freq (F) = TF (W)

3.5 Text Classification

Classification of text based on representation of featured vector. Classifications of text using K-means algorithm, Naïve Bayes and support Vector machines [6]. For mapping of text classification is possible using graphical structure and mind map structure.

4. Conclusion

Using Text mining shows the Interdependencies and co-relations between the documents can be identified which in turn would help in the appropriate grouping. In this paper on describe the main method used for text classification. The main use of text classification is to manage files in a categorized form.

References

[1] J. Jayabharathy S. Kanmani and A. AyeshaaParveen," Document Clustering and TopicDiscovery based on Semantic Similarity in Scientific Literature" in IEEE 3rd InternationalConference on Communication Software and Networks (ICCSN) , pp 425 - 429 , May 2011.
[2] Chan Wang, Caixia Yuan, Xiaojie Wang, WenweiXue," Dirichlet Process Mixture Models based Topic Identification for Short Text Streams" in IEEE7th International Conference on Natural Language Processing andKnowledge Engineering (NLP-KE), pp 80-87, Nov. 2011.
[3] Nan Liu , Yanxiang He , Qiang Chen, Min Peng and Wenqi Fang," Multi-Document Biased Summarization Based on Topic-Oriented Characteristic Database of Term-Pair Co-Occurrence" in IEEE International Conference on Information Science and Technology (ICIST), pp 832-837, March 2013
[4] Chris Clifton, Robert Cooley, Jason Rennie," TopCat : Data Mining for topic identification in a Text Corpus", IEEE transaction on Knowledge and data engineering, pp 949 -964, Aug 2004
[5] Su Yan, Xiaojun Wan ," SRRank: Leveraging Semantic Roles For Extractive Multi-Document Summarization" IEEE transaction on Audio, Speech and language,pp 2048-2058, December 2014
[6] J. Jayabharathy S. Kanmani and A. AyeshaaParveen," Document Clustering and TopicDiscovery based on Semantic Similarity in Scientific Literature" in IEEE 3rd InternationalConference on Communication Software and Networks (ICCSN) , pp 425 - 429 , May 2011.
[7] Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett," Sentence Similarity Based on Semantic Nets and Corpus Statistics" in IEEE transactions on knowledge and data engineering, pp 1138-1150, august 2006