International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Information Extraction using Tokenization And Clustering Methods

Jincymol Joseph, J R Jeba

Abstract: ***World Wide Web has become a powerful platform for retrieval and storage of information. It is a collection of text, image and multimedia files in structured, semi structured and unstructured form. These tremendous volumes of information cannot be processed so simply. An efficient and useful algorithm is required to discover information from these data. Text mining is a method for extracting meaningful information from large volume of data. Unstructured text is easily processed by humans but it is harder for machines. Text mining task involve methods such as tokenization, feature extraction and clustering.***

Keywords: Tokenization, feature extraction, cluster, stemming.

I. INTRODUCTION

World Wide Web grows rapidly and it leads to the development of irrelevant, redundant and unstructured information on the web. In order to extracting useful information from the growing data, many algorithms were developed. Web mining is an area in the field of data mining. It discovers important information present on the web pages. Web data may be in the form of structured, semi structured or unstructured format. Data extraction from structured pages is easy when compared with the other forms. Information Retrieval (IR) searches for an information in the document. It requires several pre-processing steps to structure the document and extracting features including tokenization, feature extraction and clustering. Tokenization divides the textual information into individual words. It helps to reduce the memory size of the document. Extracting sentences based on the sentence features. Sentences are classified into imperative and interrogative sentences. This classification is done based on the semantic of the sentences. Using these classifications important sentences are extracted.

II. TEXT PREPROCESSING APPROACHES

One of the key components in text mining is text preprocessing. Text preprocessing methods are tokenization, stop word removal, stemming, frequency computation, etc.

Tokenization

Tokenization is a technique in NLP(Natural Language Processing) which splits the document into tokens which are words or phrases. These tokens are used for further processing. Tokenization is useful for identifying meaningful keywords. Tokens are separated by white spaces, line breaks or punctuation marks. White spaces and punctuations do not included in the resulting tokens. It is difficult to tokenize documents without white spaces. Tokenizers provide the reliability for the documents.

Stop Word Removal

[2]Stop words are commonly used words such as a, an, the, of, etc that do not have any precise meaning on their own. A list of stop words is created and the source document is compared with this stop list. If any word matches with this list, that word is removing from the source document. Stop words do not carry any value in matching queries to the document. So, by removing the stop words did not affect the document semantics.

Stemming

Stemming is a process for reducing the modulated word to its word stem, root or base. Stemming method simply reduces a word into its base or root form. Porter stemmer algorithm is commonly used for this purpose. Stemming is a fundamental step in text mining, information retrieval and natural language processing. For example, many document contains words like prefer, preferred, preference, etc whose root word is prefer. Stemming method simply reduces these words into the root word.

III. LITERATURE REVIEW

M. Durairaj et al.[3] proposed different methods for text mining. It includes pre-processing steps such as tokenization, stop word removal and stemming. In the tokenization process, the entire document is tokenized into words or phrases. Here Nlpdotnettokenizer is used to tokenize the academic data. This tokenizer is chosen, because it is the only tokenizer that can handle large document as input and its output is easy to understand. The output from this tokenizer will be in the formatted form too. Stop word elimination is the second method for text pre-processing. Stop words are part of natural language in which it does not have much meaning. It makes the text looks heavier and has less importance. Most familiar words are articles, prepositions and pronouns. Example for stop word are in, with, a, an, etc. A dictionary based approach is used for stop word removal, where a dictionary is created with a list of English stop words. The entire document is compared with these stop words and if any match occurs, that word is removing from the array.

Manuscript published on November 30, 2019.
* Correspondence Author
Jincymol Joseph*, Assistant Professor, Department of Computer Science, St.Pius X college Rajapuram**,** Kasargod, Kerla.
Dr.J R Jeba, Associate Professor & HOD, Department of Computer Applications Noorul Islam Centre of Higher Education, Kumaracoil, Tamil Nadu.

Third technique for text pre-processing is stemming. Stemming converts words to their stems. The modified porter's algorithm is considered as a stemming algorithm which can stem with less error compared with the existing algorithm. Mohammed Safayet Arefin et.al[4] proposed method for sentence extraction. Bangla sentences are taken as input for the parsing framework. Different types of sentences with their negative forms are considered here. Then the tokenizer accepts these sentences and breaks into individual words called tokens. The token is then checked into the lexicon for validity. This framework parses the sentence with 83% accuracy. Vaishali A.Ingle[2] develop an algorithm for classifying web page information. Web page data collection is the primary step for classification process. After processing, the raw data may ends up in a data base, and this can be accessible for further processing. This raw data contains HTML tags also. By cleaning HTML coding will reduce the size of page, increase accessibility and make the search engine friendlier. Then tokenize the string into list of words. Non Negative Matrix Factorization (NMF) method is a feature extraction algorithm for producing meaningful patterns, topics or themes. Attribute – value pair is represented using the class DictVectoriz and bag-of-words approach is used for document classification based on topics. Aparna.U.R et.al[11] propose a method known as perceptron algorithm based on supervised learning. It helps to decide whether or not a feature belongs to one class or other. A weight is set based on the input. The data which are above the weight get excited will move to the next node and these data are selected. The data below the weight are rejected. Next, a truncation method is used. In this method, the selected features are truncated to the nearest possible value. So, all the features that remain closer have a common truncated value. Distance measure evaluation is done after the truncation. Features which are closer will have little difference in their distance measure. With the improved feature selection mechanism, feature extraction can be done easily. Sumya Akter et.al[5] proposed a method to extract the core content of a document. Sentence clustering approach is used for generating summary from the input document. The input document is pre-processed by tokenization and stemming operations. Term-Frequency/Inverse Document Frequency (TF/IDF) is calculated for finding word score and sentence score is calculated by summing up its constituent word score with its position. K-means clustering algorithm is used for producing the final summary of the document. A separate file is used for storing the document with its sentence score. Then the sentences are sorted in descending order. From this score, highest score is considered as centroid1 and the lowest score as centroid2. From each cluster, top k sentences are extracted and the final summary is generated.

IV.PROPOSED METHOD

In the proposed method, clustering approach is used for extracting the most important information from an input document. The document may be structured or unstructured. For content extraction, it first undergoes the noise removal process. During this process, tags such as header, footer, copyright, advertisement, etc are removing form the input document. Then for the effective summarization, it undergoes some pre-processing methods. Tokenization and stop word removal are two pre-processing methods in the proposed system. For feature extraction process, clustering approach is selected. Tokenization is a process which breaks up the input sentences into individual words or small units. It uses white spaces, punctuation marks and new line character to differentiate between two words or tokens. Porter's algorithm is a commonly using algorithm for tokenizing sentences. Disadvantage of this approach is inaccurate tokenized data. In this approach, string operations such as length of the string, string comparison, etc. applied.

Procedure Tokenization

Step:

1. Start
2. Input sentence S
3. Initialize counter and token[]
4. Find length of S
5. For each character in S
   a. Check for white spaces or new line character
      Assign words into token array
      Remove white spaces
   else
      Assign as character
6. End Loop
7. Repeat step 5 for each sentences in the input document.
8. Return token from the array token[]
9. End Procedure

Stop word removal is the second method in text pre processing. Stop words are words or symbols which does not carry any useful information to the document. Stop words makes document heavier and it is very difficult for analysts for processing. So, by removing these stop words, does not affect the meaning of the statements. For this, a list of stop word is maintained. Usually stop words are punctuations, prepositions, articles and pronouns. These are removed from the tokenized list. Removing stop words will reduce the term space dimensionality of the input document. Usually a dictionary based approach is used for this pre processing step. So, a dictionary is created with English stop words. Stop word removal phase works as follows.

1. Tokenized words are stored in an array.
2. A dictionary is created with the list of stop words.
3. Read a single stop word from the dictionary.
4. A stop word is compared with the input array and if any match occurs removes that word from the array of tokens and this comparison continues till the end of the array.
5. Another stop word is read from the stop list and continues this process until all the stop words are compared.

Here the input is tokenized text and output is stop words eliminated list.

The third method in feature extraction is clustering. In the proposed method, each input sentence is categorised as interrogative, exclamatory or imperative statements. Depends on the type of sentences clusters are defined. The organization of unlabelled data into similarity groups is called clusters.

Retrieval Number: D7943118419/2019©BEIESP
DOI:10.35940/ijrte.D7943.118419
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Procedure Feature Extraction

1. Start
2. Input Sentences for feature extraction
3. Declare array filter[], cluster1[], clustrer2[]
4. Let filter[]={?,!,please, kindly, who, when , where, how, what, ...}
5. For each sentence Si in document Di
   Compare filter[] with Si
   If found any filter[]
   Assign Si to cluster1[]
   Else
   Assign Si to cluster2[]
   Repeat until Di terminates
6. Discard Si in cluster1[]
7. Output: cluster2[] with extracted sentences

EXPERIMENTAL RESULTS

For good similarity measure, maximise intra cluster similarity and minimize inter cluster similarity. Two important similarity measures are Euclidean distance measure and Cosine similarity measure.

Euclidean distance measure,

$De(ta,tb)=\left(\sum_{t=1}^{m} | ta,-wtb|^2\right)1/2$

Where ta,tb are term weights

Cosine similarity value bounded between 0 and 1.

It is 1 if two documents are identical and 0 if two documents are dissimilar. Two external validation measures are purity and entropy. A good clustering is characterized by high purity and low entropy.

Purity= $1/N\sum_{j=1}^{L}(Wj \cap Pj)$, where Wi is set of clusters

Entropy=$1/N\sum_{j=1}^{L}(Wj \cap Ej)$

Purity is an external evaluation criterion of cluster quality.

In the existing system, clusters are formed on the basis of types of sentences present on the input document. Purity is calculated for these clusters. In the proposed method, two clusters are maintained for the input documents. Each cluster may contain different types of sentences.

Cluster 1	Cluster 2	Cluster 3	Cluster4
2	11	1	1

Purity for the existing system is 11/15=0.7333

Purity= $1/N\sum_{j=1}^{L}(Cj \cap Pj)$ is measured using this formula.

Proposed method calculates the purity based on the following table.

	Type1	Type2	Type3	Type4
Cluster 1	2	0	1	0
Cluster 2	0	12	0	0

Purity = 1/15* (2+12)= 0.93

The greater value of purity indicates good clustering. From the above analysis it is clear that the proposed method is more accurate than the existing one. Also it uses less number of clusters. So, in the above example purity reaches around 93% and it's a good sign for summarising the web documents.

V. CONCLUSION

The main objective of this proposed method is to get summarised information from the large volume of information. Proposed method uses many techniques for pre processing tasks such as noise removal, tokenization, stop word removal and clustering approach. Here, sentences are classified as interrogative, imperative and exclamatory. From this, imperative statements are selected using clustering approach. Accuracy of the proposed method reaches around 93%. In future, the sentences are analysed for paraphrasing also. Since the extracted document may contain paraphrased and repeated sentences.

REFERENCES

1. Ramalingam Sugumar & M.Rama Priya ,Improved performance of stemming using enhanced porter stemmer algorithm for information retrieval, International Journal Of Engineering Sciences & Research Technology, April 2018
2. Vaishali A , "Processing of Unstructured data for Information Extraction", Nirma University International Conference on Engineering, Nuicone-2012, 06-08 December, 2012
3. M. Durairaj, A. Alagu Karthikeyan, "Modified Porter's Algorithm for Pre-Processing Academic Feedback Data", International Journal of Pure and Applied Mathematics", Volume 118 No. 18 2018, 3009-3015
4. Mohammed Safayet Arefin, Lamia Alam, Shayla Sharmin, Mohammed Moshiul Hoque," An Empirica framework for Parsing Bangla , Assertive, Interrogative and Imperative sentences", 1st International Conference on Computer & Information Engineering, 26-27 November,2015
5. Sumya Akter, Aysa Siddika Asa, Md. Palash Uddin, Md. Delowar Hossain, Shikhor Kumer Roy, Masud Ibn Afjal, "An Extractive Text Summarization Technique for Bengali Document(s) using K-means Clustering Algorithm", IEEE 2017
6. D.S.Misbha, J.R.Jeba, "Scheduling Effective Cloud Updates in Streaming Data Warehouses using RECSS Algorithm" IJAER Vol.11 No.7
7. J.R Jeba, S.P.Victor,"A novel approach for finding item sets with hybrid strategies", International Journal of Computer Applications., Vol.17,No.5,2011
8. J.R.Jeba, S.P.Victor, "Comparison of frequent item set Mining Algorithms", International Journal of Computer Science and Information Technologies", Vol 2 (6), 2011
9. Jincymol Joseph, J.R.Jeba , "Survey on web Content Extraction" , IJAER Vol.11 No.7
10. J R Jeba, S.P.Victor, "Effective measures in Association Rule mining", International Journal of Scientific and Engineering research, Vol 3,Issue 8,2012.
11. Aparna.U.R, Shaiju Paul, "Feature Selection and Extraction in Data mining", 2016 Online International Conference on Green Engineering and Technologies (IC-GET).

AUTHORS PROFILE

Jincymol Joseph received Master of Computer Applications from Bharatiar University, Tamil Nadu, India in May 2004. She is working as Assistant Professor in the Department of Computer Science, St.Pius X College, Rajapuram. She has published papers in international journals. She is currently working towards the Ph.D. degree at the research center of Noorul Islam University, TamilNadu, India. Her area of interest is data mining

J.R.Jeba completed MCA, M.Phil and Ph.D. She is working as Associate Professor and HOD in the Department of Computer Applications ,Noorul Islam Centre for Higher Education, TamilNadu. She has published 23 international and national journals. She is a life long member of ISTE and her research area is 'Data Mining'. She is currently guiding six research scholars and two scholars completed their Ph.D under her guidance.