International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Review on Technique Used for Text and Image Categorization Using Feature Clustering

Dipak R. Pardhi[1], Charushila D. Patil[2]

[1]Assistant Professor and Head Computer Engineering Department G.C.O.E, Jalgaon,Maharashtra,India

[2]Research Scholar, M.E. (II year) Computer Engineering Department G.C.O.E Jalgaon, Maharashtra,India

Abstract: ***Many text documents having non segmented text and images needs to be divided in proper groups. There are lots of algorithms are available for classification of text and images. When text documents are classified the feature or important term available in that document is find out and depending upon that term text documents are classified. In this paper, we have to develop a similarity based Fuzzy algorithm which will categorize text document and images in to different classes by applying the concept text mining***

Keywords: Natural Language processing, Feature extraction, Concept mining, Feature clustering, Fuzzy similarity measure, Sentence level, Document level, and integrated corpora level processing.

1. Introduction

In day today life large volume of data is being handled on internet so there is a necessity of essential tool which will find, handle, filter text document .so text classification, categorization and clustering are important things which assign text

Document to predefined category according to their contents. It is important as it requires in many applications like real time email or file sorting, searching specific information, organizing data in specific groups sorting and searching of important data

Classification of text is challenging and very important field in market. A lot of research work has been done but, there is a need to divide text and images into mutually exclusive categories. We have to extract the important feature concept in the document.

Here, new fuzzy similarity based concept mining model is proposed .this model classifies set of text document in to predefined category groups. For this multiple text documents are needed to be uploaded. Every individual document is processed .each sentence from the document is separated by sentence extractor. Then syntax tress is drawn for every sentence. Here every sentence is divided in verb argument structure. Likewise every sentence is applied for syntax tree and verb argument structure is found out. Next stage is to perform word steaming .here stop words removal is done.

Then refined features are filtered from the large collection of words. Frequency of each feature is calculated in the sentence. Likewise all sentences are processed and whole document is undergone through the syntax tree and verb argument structure.
Again similar method is applied for all sentences in a document like word steaming, stop word removal and calculation of frequency.

This iterative process is applied for all the documents. And finally refined feature with their actual frequency is calculated.[1]

Next step is to apply classes to the documents Class is nothing but the no. of clusters /groups in which the text documents are to be divided. [4] Depending upon the similarity of features available in documents text documents are classified in classes.
Due to this method feature reduction is possible and high performance is achieved. This model uses FFCSA which checks each extracted feature to avoid duplication and achieves ambiguity removal.

A. **Natural Language Processing**:
Natural Language Processing is main branch of artificial intelligence and classification of text is important area in which each text documents is processed by finding out their grammatical syntax and semantics. Text mining contains two basic techniques:
i Text classification
ii Text clustering

i. Text Classification
It is important method which text documents are divided in to various categories using supervises learning technique. There are large number of text categorization techniques are available but the main problem of all is. They generate large set of features. When categorization algorithm is used large volume of feature sets are created which requires more memory space and time. To avoid this, new technique must be developed which will compress resultant features as well as original meaning of them will not lost and high performance should be achieved.

ii. Text Clustering
It is one of the traditional method which uses unsupervised learning paradigm. It divides document into groups or categories where each group represent some topic that is different than another groups.[5]. In Proposed method, text documents are processed on sentence level, then document and lastly integrated corpora level. At each level, snetences is processed features are extracted and matrix of features is created thend features are reduced, edundant entries are removed .the process continues from lowest i.e. sentence level up to highest corpora level. Finally, resultant refined matrix is used to make Support Vector machine Classifier.

iii. Fuzzy Logic

It is a mathematical model in which truth can be partial i.e. result is not fix true and false but it has it can have value between 0 and 1, that is not completely false and completely true [6]. It is based on approximate reasoning instead of exact reasoning.

2. Related Work

In [1] Marcus Vinicius C. Guelpeli Ana Cristina Bicharra Garcia proposed a text categorizer using the methodology of Fuzzy Similarity. Where the grouping algorithms Stars and Cliques are adopted in the Agglomerative Hierarchical method and they identify the groups of texts by specifying some time of relationship rule to create categories based on the similarity analysis of the textual terms.

The proposal is that based on the methodology suggested, categories can be created from the analysis of the degree of similarity of the texts to be classified, without needing to determine the number of initial categories. The combination of techniques proposed in the categorizer's phases brought satisfactory results, proving to be efficient in textual classification.

Thus their work proposed a text categorization based on the Agglomerative Hierarchical methodology with the use of fuzzy logic.

In [2] the Author L Choochart Haruechaiyasak, Mei-Ling Shyu suggested a method of automatically classifying Web documents into a set of categories using the fuzzy association concept is proposed. To solve ambiguity problem, fuzzy association is used to capture the relationships among different index terms or keywords in the documents i.e., each pair of words has an associated value to distinguish itself from the others. Therefore, the ambiguity in word usage is avoided. . The analysis of results show that their approach yields higher accuracy compared to the vector space model

In [4] the Author: Ahmad T. Al-Taani, and Noor Aldeen K. Al-Awad suggested a fuzzy similarity approach for Arabic web pages classification is presented. The approach uses a fuzzy term-category relation by manipulating membership degree for the training data and the degree value for a test web page The approach used fuzzy term-category relation by manipulating membership degree for the training data and the degree value for a test web page. They used and compared six measures in this study. These measures are: Einstein, Hamacher, bounded difference, Algebraic, MinMax, and Special case fuzzy (Scfuzzy). The best performance is achieved by the Einstein measure then the Bounded measure followed by Algebraic measure.

The training data is first collected from different sources, and then normalized by passing it through the noise elimination module. The approach also includes the HTML stripping, stop word removing, and stemming. The learning process began by representing terms as numbers to reduce their representation. The final step in the process was to apply the six measures to the web pages.

In [5] the author Shady Shehata, and Fakhri Karray suggested a new concept based mining model composed of four components to improve the text clustering quality. The first component is the sentence-based concept analysis which analyzes the semantic structure of each sentence to capture the sentence concepts using the proposed conceptual term frequency ctf measure. Then, the second component, document-based concept analysis, analyzes each concept at the document level using the concept-based term frequency tf. The third component analyzes concepts on the corpus level using the document frequency df global measure. The fourth component is the concept-based similarity measure which allows measuring the importance of each concept with respect to the semantics of the sentence, the topic of the document, and the discrimination among documents in a corpus. By combining the factors affecting the weights of concepts on the sentence, document, and corpus levels, a concept-based similarity measure that is capable of the accurate calculation of `concept matching and concept-based similarity calculations among documents in a very robust and accurate way. The quality of text clustering achieved by this model significantly surpasses the traditional single term-based approaches. In[3] the author Shalini Puri and Sona Kaushik discussed different fuzzy similarity related algorithms and methodologies in detail. Which generates good results with the underlying techniques, mechanisms and methodologies?

These models focus on new kinds of different classification issues and techniques. And contribute in providing the information about advanced fuzzy classification, related models and techniques. The analytical review provides a simple summary of the sources in an organizational pattern and Combines both summary and synthesis to give a new interpretation of old material. Additionally, their experimental results and their parametric data are sufficiently described and compared independently. Such comparative studied and technical analysis charts provide a strong base to understand the use of fuzzy and its related concerns. Various experimental results have proven themselves good for the models and techniques. The utility of fuzzy logic and its areas give a good effect on text mining and text classification.

Three important concepts are used related to text document classification

1) Terms Used In Fuzzy Model:

A. Concept mining

Text document contain many words including articles, nouns, adverbs etc. this model is used to extract the concept which is embedded in document. Here concept means word/ feature which have proper semantic structure on sentence.

B. Feature extraction

As concept mining finds out important concept in the document, the concept or feature is extracted form document as well as reduced mean duplicate entries are avoided.

C. Similarity Measure

After concept mining and feature extraction the important thing is similarity measure among text documents. This similarity is not exact so fuzzy logic is applied here.as fuzziness gives uncertainty and provides range of values.so two documents having maximum similar features are

grouped in one category for this various fuzzy algorithms are applied like fuzzy c-means, means etc,.

2) Proposed Work

Text document categorization is done up till now. In the proposed work we have to implement new similarity based fuzzy model in which fuzzy algorithm is implemented which will categorize images depending upon text associated with it. In addition to text document new section is added i.e. image with text where each image is associated with a text box. The text related to image contains meaningful information about the image like contents of that image.

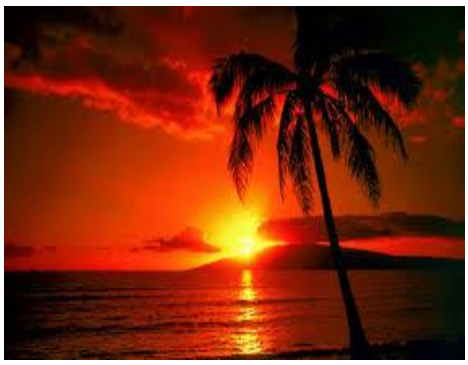

Figure 1: Image 1

While uploading above fig. user will upload the contents belonging to that image in the text box associated with this image. Like sun set, coconut tree sea beach are the actual contents in the image. In future when user will search the image the image will be retrieved not by its name but by its content in the image.

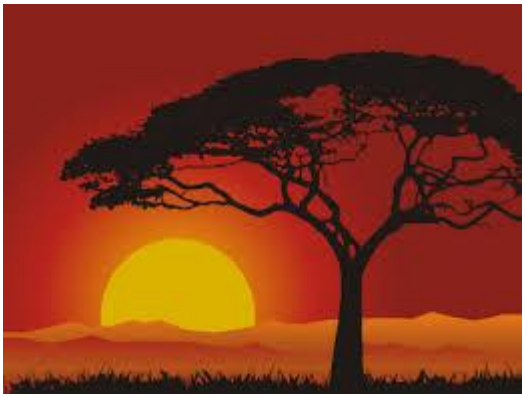

Figure 2: image 2

Likewise when above image is uploaded, the text box will contain the features sun set and tree. Thus images are saved by loading the content or important feature in it. Our Fuzzy based text mining model will work in the same manner like that of text categorization. Thus each text document is scanned and sentence level processing is done. Word steaming, stop words removal stages are carried out. Feature vector is created. Then refined feature vector is created. And redundant entries are removed. This process is carried out for all images with text and image categorization is done .thus images with similar features are grouped in one cluster while different images are in another cluster.

3. Conclusion

In this paper, we have implemented image categorization using text mining and feature clustering. Where English language was used. In future, we can use this technics for another language also.

References

[1] Marcus Vinicius C. Guelpeli Ana Cristina Bicharra, "Constructed Categories for Textual Classification Using Fuzzy Similarity and Agglomerative Hierarchical Methods"
[2] L Choochart, Web Document Classification Based on Fuzzy Association"
[3] Shalini Puri1 and Sona Kaushik. "A Technical Study And Analysis On Fuzzy Similarity Based Models For Text Classification"
[4] Ahmad T. Al-Taani, and Noor Aldeen K. Al-Awad "A Comparative Study of Web-pages Classification Methods using Fuzzy Operators Applied to Arabic Web-pages"
[5] Shady Shehata, and Fakhri Karray, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering"