International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 4, Issue 3, May-June 2015 ISSN 2278-6856

Image Classification Using Text Mining and Feature Clustering

(Text Document and Image Categorization Using Fuzzy Similarity Based Feature Clustering)

[1]Mr. Dipak R. Pardhi , [2]Mrs. Charushila D. Pati

[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

In traditional text mining many text documents are separated and clustered by considering similarity feature among text document. Various classification algorithms are used for text categorization, but image separation is not available considering text mining concept. In this paper, we have developed image clustering and categorization technics by using the concept of text mining. Here multiple images are uploaded .every images has given some meaningful text considering the contents in it. And our fuzzy algorithm is applied to images and images are separated by applying text mining techniques.

Keywords: Feature extraction, Concept mining, Feature clustering, sentence extractor

1. INTRODUCTION

On day today life many text documents have to be processed as there is need of text categorization system. Text classification is challenging and very important field in market. A lot of research work has been done but, there is a need to categories a collection of images by the technics of text mining.[7] Thus we have to categories a collection of images into different classes. Here For this, multiple images are uploaded and while uploading each image, one has to extract feature/concept by using supervised learning method and different text classification algorithm to be used.

In this paper, we have applied fuzzy logic to text mining for searching and clustering of image into selected groups of clusters. Here, user can upload multiple images.

While uploading each image specific valid information regarding that image has be given which will act like annotation for that image. .

Every image has assigned a text box in which user will upload information related to that image like description in brief about that image considering content in that image. The sentence extractor will separate each sentence in that description. Thus each sentence is extracted by sentence extractor and spitted . Next step is to remove stop words and perform word steaming.

Then important feature from each sentence is extracted. Frequency is calculated of those features. The duplicate words are removed and redundant entries are updated.

This process is applied to all sentences in one document of image.

The same is done for all images. And reduced feature vectors are calculated.

At the last stage, user has to assign classes to these images. By considering common features among all images. Thus similar type of images is categorized in one group and different type of images is by default enters in different groups. i.e. clustering is done.

Natural language processing is important branch of artificial intelligence and text classification is important area where text documents are processed by finding out their grammatical syntax and semantics.[1]

Text mining contains two basic techniques:

Text Mining is a new, multidisciplinary field, which includes spheres of knowledge like Computing, Statistics, Linguistics and Cognitive Science. It include extracting regularities, patterns or trends in large volumes of texts written in a natural language,

It is inspired by Data Mining. [3]

It can be applied in a variety of contexts:

In the creation of summaries like cauterization, feature extraction, text categorization.

Basically it contains two methods

i Text classification

ii Text clustering

A. Text classification

It is important method which categorizes text document using supervise learning technique. Various text categorization techniques are available but most of them face a big problem of 'Curse of Dimensionality'. When categorization algorithm is used large volume of feature sets are created. Due to which more memory space and time is required.[4]

To avoid this, feature set must be reduced. as well as original meaning should not lost and high performance should be achieved.

B. Text clustering

It is useful for performing text classification.[5]

In this paper, we propose a fuzzy similarity-based model for feature clustering.

It attempt to divides text document into sentences. The feature from each sentence is extracted.

Thus At each level, features are extracted and feature matrix is created. Then total number of counted features

is analyzed. If repeated features are found, then those duplicate entries are removed. And frequencies of features are updated. Each sentence is processed and whole document ids processed according to this. Thus important feature from each document represent that image .further it helps to perform clustering.

2. RELATED WORK

Marcus Vinicius C. Guelpeli et,el [1] 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.

Author L Choochart et.el[2] 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

Author: Ahmad T. Al-Taani,et. el[4] 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.

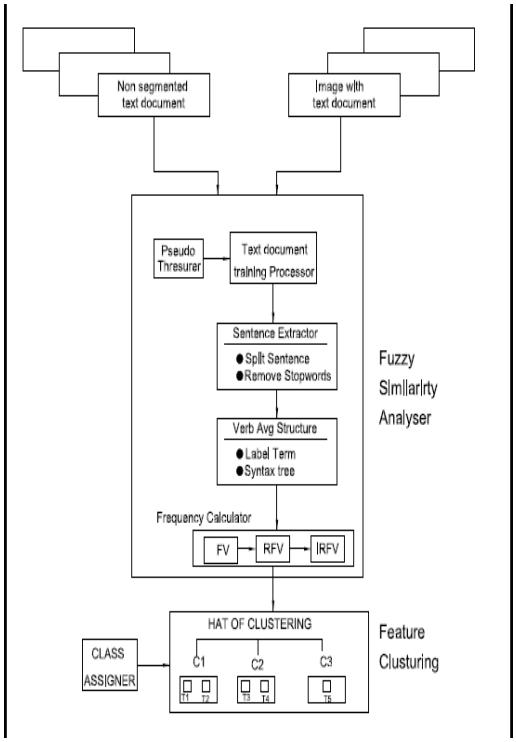
Author Shady Shehata, and Fakhri Karray et. el[5] suggested a new concept based mining model composed of four components to improve the text clustering quality. The first component is the sentence-based concept analysis which analyzes the semantic structure of each sentence to capture the sentence concepts 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 pair wise documents is devised. This allows performing 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.

The author Shalini Puri and Sona Kaushik[3] 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 [3]. 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. [4]

3. PROPSED WORK

A Architecture Of Proposed Work



The important concepts are used related to text document classification are as follows:

Fuzzy Logic

It a mathematical logic whose answer can be between 0 and 1 i.e. neither fully true nor false thus it provides approximate reasoning instead of exact.[5]

Clustering

By using fuzzy logic and concept of text mining, we can divide similar type of image into different classes thus we have used fuzzy algorithm to cluster images.[8]

This algorithm calculate frequency of every feature and remove redundant entries according to common feature, images are grouped into classes

Concept mining

It is used to extract the concept which is embedded in text document. Concept means word/ feature which have proper semantic structure on sentence.

Feature extraction

In text classification, large number of feature is generated. Due to which speed of execution is minimized and storage requirement is increased. So the feature reduction is important method which is to be done.

There are two methods of feature reduction

a) Feature extraction
b) Feature selection

In feature extraction, original feature set is converted in to different reduced feature set. Which is better method than feature selection?

3) Similarity Measure

Fuzziness gives uncertainty and provides range of values. Fuzzy similarity measurement is used to find out similarity among different text document, so similar type of text document can be grouped in to one category group.

B. Problem Definition:

To classify multiple images in to predefined classes and perform feature clustering using the content of images.

In this paper, the description related image is considered to be the combination of words.

Fuzzy logic is applied to each image to cluster it in different classes.

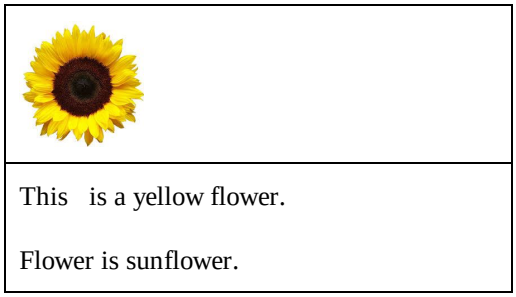
C. Case Study

Consider three images Img1, Img2, Img3.

Each image is associated with text document in which valid brief information related to that image is given which is combination various sentences.

Each text document is processed through the sentence, document and integrated corpora levels.

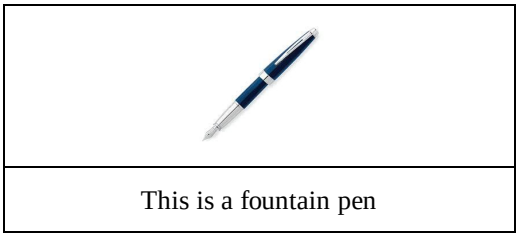
Consider following 3 images along with their text attached to it.



This is a yellow flower.

Flower is sunflower.

Img1



This is a fountain pen

Img2

Rose flower is good

Img3

Our algorithm will work as follows:

S1: This is a yellow flower.

S2: Flower is sunflower.

1Word stemming and stop word removal:

S1:

Feature no.	Features	Frequency
F11	This	1
F12	is	1
F13	a	1
F14	yellow	1
F15	flower	1

S2:

Feature no.	Features	Frequency
F21	flower	1
F22	is	1
F23	Sunflower	1

Here a, this, it, is stop words so they are removed.
So refined features of both sentences are:

Feature no.	Features	Frequency
F14	yellow	1
F15	flower	1

S2:

Feature no.	Features	Frequency
F21	flower	1
F23	Sunflower	1

2. Removal of redundant entries and combine both sentences:

Feature no.	Features	Frequency
F1	yellow	1
F2	flower	2
F3	Sunflower	1

Likewise all documents corresponding to images are processed.

Features from other sentences are extracted and processed as follows:
S2: fountain, pen.
S3: rose, flower

Integrated features (for all images):

Feature no.	Features	Img 1	Img 2	Img 3
F1	fountain	0	1	0
F2	**flower**	**2**	**0**	**1**
F3	pen	0	1	0
F4	Sunflower	1	0	0
F5	yellow	1	0	0

Classes would be two: C1 and C2. And Img1 and Img2 have at least one common feature so images are clustered in 2 classes like:

1) C1: Img1 and Img2
2) C2: Img2

D. Algorithm

1. Uploading of images
2. Give brief description to each image
3. Preprocessing of images

a) Sentence level processing by sentence extractor. Separate each sentence
b) Draw syntax tress divide each sentence in verb argument structure.
c). calculate frequency of each feature.
d) Perform word steaming and remove stop words.
f) Remove redundant entries update their frequency

Follow step a to f for all sentences in one text document related to image.
Follow step 3 for all text document in diff images,

4. Update frequency for all images
5. Define classes in which images are to be clustered
6. Depending upon common features divide images into categories

Then Mathematical term for image classification –
Consider images i1,i2,i3...in to be classified among classes c1,c2,c3…cn
Where each image I is associated with text box t.
Where i1€t1, i2 i€t1, i3€t3 and tn€tn

$$T_B = \{ t_1, t_2, t_3, \ldots\ldots t_n \} \quad (1)$$

At Sentence level processing:

A text t_1 is composed of set of 'n'

$$t_1 = \{ S_{i1}, S_{i2}, \ldots\ldots S_{im} \} \quad (2)$$

where i = text document no
m = no of sentences
$m = \text{size of } \{T_1\}$

Feature Vector (FV)

$$FV = \{F_{i1}, F_{i2}, \ldots\ldots F_{in}\} \quad 3$$

Reduced Feature Vector (RFV)

$$RFV = \{RFV_{i1}, RFV_{i2}, \ldots\ldots RFV_{in},\} \quad 4$$

At integrated corpora level

Suppose classes =

$$Classes = C_1, C_2, \ldots\ldots C_n\} \text{ and}$$

$$Classes = C_1, C_2, C_3.$$

And

$$Images = I_1, I_2, I_3, I_4,$$

Then

$$C_1 = \{I_1, I_2,\}$$
$$C_2 = \{I_3\}$$
$$C_3 = \{I_4\}$$

$$\text{Where } C_1 = Max\,(\Sigma(RFV)) \quad 5$$

$$\text{Let } Max\,(\Sigma(RFV)) = \delta$$

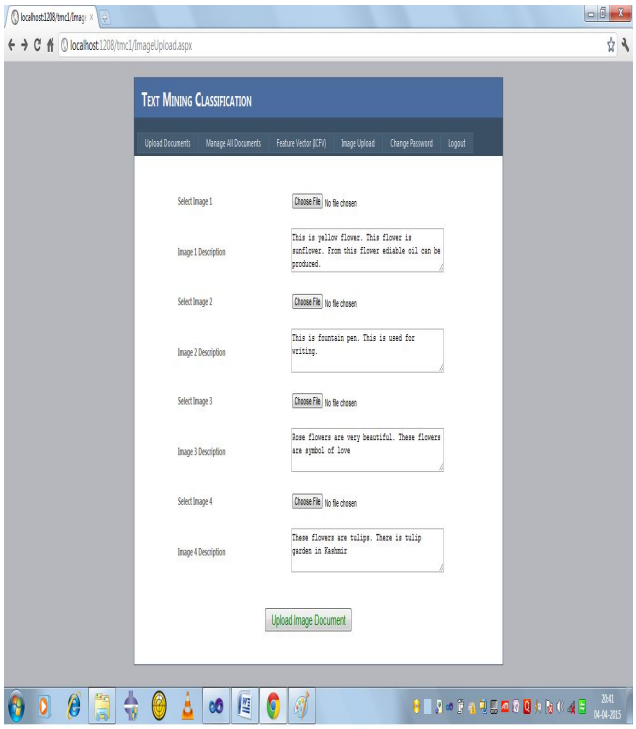
Which is threshold frequency for all images for classification

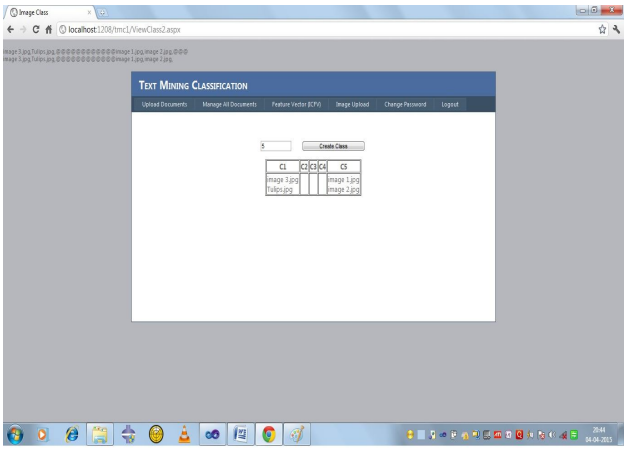
$$C_2 = \log_{P1}^{\delta\text{-}1}(\Sigma(RFV))$$

$$_{1}C_3 = \log_{P1\text{-}}(\Sigma(RFV))$$

$$\text{Where } P_1 = \frac{\delta}{\text{NO OF CLASSSES}}$$

5. RESULT





6. CONCLUSION AND FUTURE SCOPE

Thus fuzzy similarity algorithm of text mining can be applied for making clusters of various images. Thus image categorization is done by using the technic of text mining .In future fuzzy algorithm can also be applied for audio, video data set classification.

REFERANCES

[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"
[6]. Shalini Puri and Sona Kaushik "An Enhanced Fuzzy Similarity Based Concept-Based Mining Model for Enhancing Text Clustering"
[7]. Shalini Puri and Sona Kaushik, "Sensitive Information on Move"
[8]. Fathi H. Saad1, Omer I. E. Mohamed2, and Rafa E. Al-Qutaish2, "comparison of hierarchical agglomerative Algorithms for clustering medical Documents