International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 06 | June-2016 www.irjet.net p-ISSN: 2395-0072

Web Document Clustering System Using Fuzzy Logic and Feature Extraction

Pranali Raut[1], Prof. Nilesh Khochare[2]

[1]Student, Department computer, JSPM NTC Pune, Maharashtra, India
2Professor, Department computer, JSPM NTC Pune, Maharashtra, India

---

Abstract - *In computer analysis, many files are usually analysis much of the data but in those files consists of unstructured data, and that data examined by computer examiners are more difficult to be performed. So, it need the automatic technique to analysis that unstructured documents. For this purpose it uses clustering documents, it can gives new and useful knowledge from the documents under analysis. Clustering technique is used to help grouping of documents which are closely related to each other. Recently used clustering algorithms have some disadvantage like data preparation, outliers and high sensitivity. As a solution of above problem develop the new clustering technique and the main theme is web documents is converted into clustering documents with the help of data prepossessing, features extraction, and weighted scores matrix techniques.*

Key Words: **Web Documents, Web Crawler, Fuzzy Logic, Weighted score Matrix, Feature extraction, Clustered Document.**

1. INTRODUCTION

Web documents are complex and heterogeneous and the linking of web pages is very difficult and complex in nature. The simple idea of web clustering is hundreds of thousands of files are usually examined to come to a conclusion. So, there is a need to discover the fast method that can group the required documents. Web documents contain many data which is in not structured format so that data is analysis by any machine is not possible. So, automated methods of document examined are of great interest.

Natural language processing (NLP) means the automatic processing of the human language. Some NLP applications are grammar and spelling checking, optical character recognition, information retrieval, document classification, clustering and information extraction. It is an area of computer science and artificial intelligence concerned with the interactions between computers and human natural languages. Deriving meaning from human or natural language input tends to a challenging factor in NLP which requires natural language understanding. Natural language processing has ability to process sentences in natural language like English, rather than computer language such as C++.

In data mining process, data pre-processing is most important step. In data pre-processing, it will removes irrelevant, noisy and unreliable data so output of data is good quality of data. It converts raw data into an understandable format with the help of cleaning, transformation, selection and feature extraction. In data cleaning, it fills in missing values, identify outliers and remove them, remove noisy data. In data transformation, aggregation and normalization takes place. In data reduction, it reduces the volume but it produces the similar and equivalent analytically results and data discretization is part of data reduction as well as it replace the numerical attributes with nominal once.

Representation of text is useful for the selecting features to represent text that will be clustered. Feature selection is a process of identifying the most effective subset of the original features to be used in clustering.

Feature extraction mainly reduces the amount of resources required to describe a huge set of data. Analysis of big number of variables it requires a large amount of memory as well as computation power which generalizes poorly to new samples. Extraction of features is the process of using linear or non-linear transformations on original features to generate projected features to be used in clustering. It is used for methods of combination of the variables to get around these problems while still describing the data with more accuracy.

Internet has becomes the huge data repository, so that is face some problem of information overload. Many users use the World Wide Web for taking or getting the required information. It contains complex and dynamic nature of the Web; it is a tedious process for the average user for information retrieval. Search engines, Web Directories and meta-search engines have been developed in order to help the users fastly and easily satisfy their required information.

So it need to development of new techniques which gives ultimate goal of finding best matching of their needs. One technique that can achieve an important role to this objective is document clustering.

This paper can be classified as follows: Section 1 dedicated for Introduction. Section 2 reserved for Literature Survey, Section 3 is allocated for Proposed System. Section 4 is dedicated for Results discussion and finally section 5 is done with conclusion.

2. LITERATURE SURVEY

To put forward the idea of "Web document clustering system using Fuzzy Logic and Feature extraction". This paper analyzes many concepts of different authors as mentioned below:

Khaled M. Fouad narrates that volume of information is growing continually; there is increases interest in helping people good to find, filter and manage these resources. Text clustering is the process of grouping documents that have closely related properties based on semantic and statistical content, is an important component in many Information retrieval. The proposed agent aims at providing qualitative improvement over traditional VSM by using semantic based model based on Word-net ontology. The newly developed semantic based model layout gives to increased performance and get a clustering be efficient. It also eliminated the problems existing in the VSM commonly used for clustering. The clustering result based on semantic model has more efficiency values and faster than those based on the traditional VSM.

Dr.T.Nalini explains clustering means the process of grouping of data and that grouping is done by searching semantics between data related to their characteristics. A study of clustering algorithms across two different data items is evaluated here. Which clustering algorithm is best is decided by comparing with one or more clustering algorithms based on the time taken to form the estimated clusters. The experimental results of many clustering algorithms to make clusters are depicted as a graph. So it can be concluding as the time taken to form the clusters increases as the number of cluster increases. The simple K-Means takes the longest time to perform clustering. It has some disadvantages like it does not identify outliers and not suitable for different size of cluster.

George Forman describes more research in speed up text mining involves algorithmic improvement for many large scale applications, like classifying huge document repositories. Feature extraction is widely used to increase the efficiency of document clustering algorithm. The time takes in extracting word features from texts can itself greatly increase the initial training time. Quality of cluster to be form is totally depends on the quality of the extracted features. This introduce a fast method for text feature extraction that folds together string hash computation, Unicode conversion, word boundary detection and forced lower casing. It show empirically that integer hash features result in classifiers with equivalent statistical performance to those built using string word features, but they require adequate computation as well as few amount of memory. Speedy FX has some methods which previously used for extraction of features. Using speedy FX integer hashes it runs faster, required less memory for transmission and use multiple classifiers and has an effect on classification performance.

Giridhar N S explains Information Retrieval System retrieves the document from huge documents of collection. This is need by the people and it should fulfill the users need also. IR System applied Stemming algorithms in the preprocessing stage to get the words to their root form. So increases the retrieval performance of the system. More such stemming algorithms exist; from those porters stemming algorithm is the famous one because of its simplicity, efficiency and availability. But has some drawbacks like output stems are meaningless. It is suitable for American English but people follow British English. In this a TWIG produces meaningful stem and it reduces the error rate. But it should improve efficiency and simplicity of TWIG algorithm.

3. PROPOSED SYSTEM

The main theme is Web Documents is converted into Clustering Documents with the help of Data Preprocessing, Features Extraction and Weighted Scores Matrix Techniques.

3.1 System Overview:

The proposed method of "Web document clustering system using Fuzzy Logic and Feature extraction" Can be described efficiently according to the steps which are depicted in below figure:

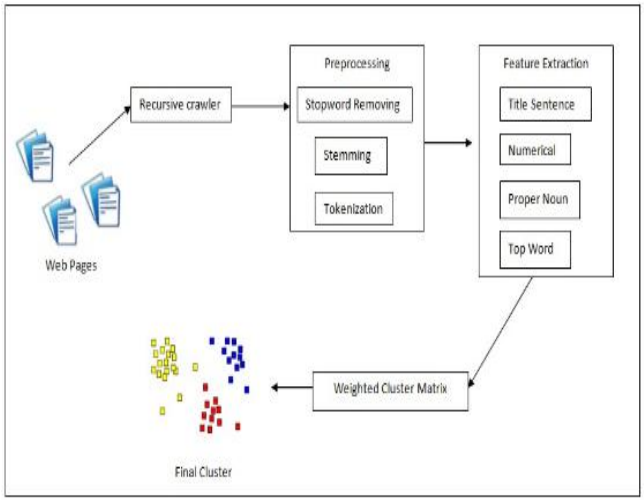

Fig -1: Overall System Diagram.

Step1: This system first creates an interactive web crawler which eventually parses the web pages and collects the data and saves in .txt file format. Then the folder in which these web data is stored is given as the input to the system.

Step2: After giving input to the system then next is data preprocessing takes place. In data preprocessing contains:

a. Removing special symbol: It contains removing of unwanted symbols from that data. For example: ? ,]

b. Removing stop word: stop word usually refer to the most common words in a language. These words not convey any meaning. Removing these words will save spaces for storing documents. Example: is, a, the, an, for, all

c. Stemming technique: stem is common root which form words with the similar meaning but appear in many morphological forms. Such as faster, fastest and fast has same meaning and root word is fast.

Step3: After process of data preprocessing, data is becomes pure. After that next step is feature extraction. In that,

a. Title Sentences

b. Numeric Words

c. Proper Nouns

d. Term Weights

Feature extraction method is widely used to increase the efficiency of document clustering algorithm. So quality of cluster to be form is totally depends on the quality of the extracted features. User dealing with document clustering a care should be taken while selecting extraction technique.

Step4: In step3 generate output, that output gives input to the weighted matrix score. Weighted score matrix used to specify of importance of criteria level. Assigning meaning to weighting factors is subjective. For this reasons, keep the number of weighting factors small. In weighted matrix score, system creates a score matrix of all the documents by comparing with one another to yield a score matrix which contains aggregate feature score. The grouping of these values represents the most accurate clustered documents.

Step5: When matrix is created then fuzzy logic technique is used. In fuzzy logic, taking the decision based on weighted score matrix.

Step6: On the basis of fuzzy logic clustering documents will be done. Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.

In this way web documents is converted into document clustering.

3.2 Algorithms:

- Preprocessing Algorithm:

Step 0: Start

Step 1: Read string

Step 2: divide string into words on space and store in a vector V

Step 3: Remove Special Symbols

Step 4: Identify Stopwords

Step 5: Remove Stopwords

Step 6: Identify Stemming Substring

Step 7: Replace Substring to desire String

Step 8: Concatenate Strings

Step 9: stop

This preprocessing algorithm is use for removing special symbols, Stopwords and for stemming.

- Algorithm for Document Clustering Using Fuzzy Matrix Weighted Method:

Input: Merged Feature vector Fv User Accuracy as Ua

Output: Cluster Set C= c1, c2, c3.cn

Step 0: start

Step 1: create matrix M of length Fv

Step 1: For i=0 to Fv length (for each row)

Step 2: For j=0 to Fv length (for each column)

Step 3: Fvr= element of one row

Step 4: Fvc=element of one column

Step 5: Compare features and get score as Sc

Step 6: Average Score as Asc=Sc/4

Step 7: add Average to matrix M

Step 8: End Inner For

Step 9: End Outer For

Step 10: for every file in Ms Rows if (Asc less than or equal to Ua) then add into cluster Ci

Step 11: return cluster set C

Step 12: Stop

- Algorithm To Find Noun:

Step 0: Start

Step 1: Read string

Step 2: divide string into words on space and store in a vector V

Step 3: Identify the duplicate words in the vector and remove them

Step 4: for i=0 to N (Where N is length of V)

Step 5: for ith word of N check for its occurrence in Dictionary

Step 6: if present then return true

Step 7: else return false

Step 8: stop

3.3 Mathematical Model:

	Mathematical Model	Observations
1.	$f(T_{100}) = \int_0^t Tw \in D$ where D=document	Calculating top 100 words
2.	$f(N_D) = Tw \in N_D$ where $N_D \Rightarrow$ numerical data	Deciding the sentence which contain numerical data
3.	$f(V_c) = \int f_{(T_{100})+f(T_t)+f(N_D)}$	Formation of cluster vector
4.	$f(R_{T100}) \sum_{=0}^{N} Si \in f(T100)$	Deciding rank of top 100
5.	$f(R_t) = \sum_{0}^{N} Si \in f(Tt)$	Deciding rank of Title words
6.	$f(R_{ND}) = \sum_{0}^{N} Si \in f(ND)$	Ranking of numerical data
7.	$f(S_V) = f(R_{T100}) + f(R_{Tt}) + f(R_{ND})$	Merging that is searching the words
8.	$f(D_c) = \sum_{0}^{K} f(Sv) \in di$	Classification on documents, that is Cluster the documents

4. RESULT AND DISCUSSION

To show the effectiveness of proposed system some experiments are conducted on java based windows machine using NetBeans as IDE and evaluate the system based on following graphs:

- Clustering performance for different no of file:

Fig -2: Clustering performance for different no of files.

The plot figure 2 drawn for different no of files to cluster. The plot Indicate less time for clustering more number of files. This expresses our clustering time is not directly proportional to number of files, so it can be conclude the best method of clustering unstructured documents.

- Average Cluster creation for Given Documents:

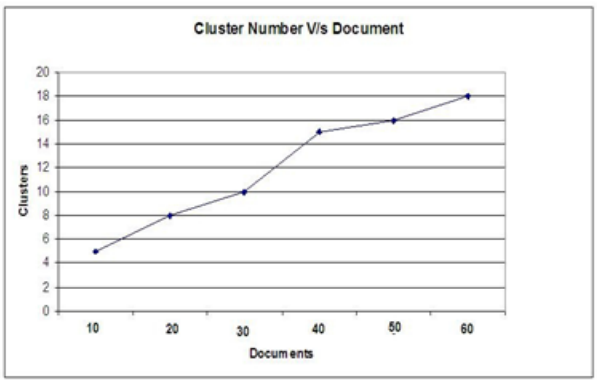

Fig -3: Average Cluster creation for Given Documents.

The plot figure 3 drawn to show number of cluster can create on increasing amount of documents. The warm indicates steadily increasing in number of clusters on increasing of documents. It shows our methodology proportionate to number of documents, so it over performs clustering technique.

5. CONCLUSION

This paper successfully accumulates most of the techniques of many authors as described in section 2 of related work. So, by analyzing all methods it seem to be like number of method is perfect in providing solution for "Web document clustering system using Fuzzy Logic and Feature extraction". As an effort to this, this paper tries to improve the concept of "Web document clustering system using Fuzzy Logic and Feature extraction" by introducing clustering based techniques is to extract features from the web documents.

REFERENCES

[1] M. Elsner, E. Charniak, and M. Johnson, "Structured generative models for unsupervised named-entity clustering," in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 09). Stroundsburg, PA: Association for Computational Linguistics, 2009, pp. 164172.

[2] V. Loia,W. Pedrycz, and S. Senatore, "Semantic web content analysis: A study in proximity-based collaborative clustering," IEEE T. Fuzzy Systems, vol. 15, no. 6, pp. 12941312, 2007.

[3] H. L. Larsen, "An approach to flexible information access systems using soft computing," in Proc. of the 32nd Annual Hawaii International Conference on System Sciences, Hawaii, 1999, p. 231.

[4] W. B. Frakes and R. Baeza-Yates, Information Retrieval Data Structures and Algorithms. Englewood Cliffs, NJ: Prentice Hall, 1992.

[5] S. Park, D. U. An, B. R. Cha, and C. W. Kim, "Document clustering with cluster refinement and non-negative matrix factorization," in Proceedings of the 16th International Conference on Neural Information Processing, Bangkok, Thailand, 2009, pp. 281288.

[6] T. Kohonen, Self-Organization Maps. Berlin Heidelberg: Springer- Verlag, 1995.

[7] J. MacQueen, "Isome methods for classification and analysis of multivariate observations," in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1. Berkeley, CA: University of California Press, 1967, pp. 281297.

[8] K. Jain and R. C. Dubes, Algorithms for Clustering Data. Prentice Hall, 1988.

[9] S. Lu and K. Fu, "A sentence-to-sentence clustering procedure for pattern analysis," IEEE Transactions on Systems, Man and Cybernetics, vol. 8, pp. 381389, 1978.

[10] O. Zamir and O. Etzioni, "Web document clustering: a feasibility demonstration," in Proceedings of 19th international ACM SIGIR conference on research and development in information retrieval (SIGIR 98), 1998, pp. 4654.