# INFORMATION RETRIEVAL USING TEXT LEVEL CLUSTERING USING PAGE RANKING TECHNIQUE

D.Saravanan.

Faculty of operations & IT, IFHE University, IBS Hyderabad, India.

**Abstract**

In the aspects of mining, it is used to extract the data's in the efficient manner and then fast retrieval of data's. The current aspect is to clustering the sentence level text by using the proposed algorithm. By means of that, it allows patterns to all clusters .we give a text in a sentence or a sentence that has to be relatively present in documents or a set of documents. Every clustering algorithm finds the association between the data points. Based on the closeness between the data points clustering are take place. Item are very close each other they are all in one group. Items are not similar they are all form in other groups Techniques are finding similarities between the given text. Based on data and the data objects by using the novel fuzzy clustering pattern.

*Key terms: Information extraction, Clustering, Spectral grouping, Similarity indexing, Outlier.*

## I. INTRODUCTION

**C**lustering is the technique form the similar items in one groups, unsimilar items are in other groups [1]. Based on the input data forming cluster are vary, if the data set are consisting of pure data without any noise and repeated data, clustering output also very high. Cluster can have applied to every field today it helps to form the group between the item sets. Technique applied image, text and other data type [2]. In text data mining it brings similar text are in one group. Different techniques are used to group the similar text. Text mining and clustering both output is same, both techniques are forming the similar text in one group. Vector match technique are used to forming the clustering of text data, this technique is performing the clustering based on the similarity based on the give input text data set. After forming groups cluster process continues finds similarity between the groups. Text mining techniques are different from other type of clustering, technique based on the sentence length formation of cluster are get differed based on the available text length. Another major challenge in this text mining, language difference based on the language this formation get differ. For small phrases this formation gets varied. In some language one sentence gives the different meaning, it gives challenge to the researchers and developers. This formation used for many text mining applications in day to life. Most of the search engines are used this technique to find user most relevant information Mining are take place based on the group or categories or topics based on that text classification are done effectively. Today most of the developers used this technique for classify the content and also improve the users searching functionality. This technique used for other application such as construction the decision trees based on the division, categorization of information. Hierarchical structures are constructed very effectively by using this. Another most important application of text mining is rule based operation. Based on the rules information are classified, this application are widely used in banking sector to find the credit balance to the customer also to sanction any loans are done by this. Based on rule based classification find easily particular customer are eligible or not for certain loan process.

## II. LITERATURE REVIEW

Pedrycz, et al., [3] proposed the new idea in fuzzy clustering technique, technique used for unsupervised learning methodology. Fuzzy based type of clustering is used for grouping the image, pattern type of data sets effectively. Here based on the similarities between the data points are consider for clustering the information. If data points similarities are more than they are all grouped into one cluster, followed by next higher similarity like cluster formations are done. Clustering also done with labeled grouping. Because of this tech-nique clustering performance get improved this proved in the experimental results. Experimental outcomes also veryfied that proposed technique efficiency get improved than the existing techniques.

Yuhua et al., [4] proposed idea about text based techniques using sentence based uniqueness. Existing technique are based on the log based approach this give long sentence of information, this bring the time complexity. This paper reduces the problem find in the existing technique, technique used in this paper using lexical data base model. This bring instead of using long sentence, information is used small text. With help of shorter text finding the similarities between the text are evaluated. Testing are done with help of two pair of sentence are selected. Entire technique used for short text of sentences. Because of this input techniques are implemented in variety of applications. Every short or long sentence are made of words, based on this word comparison technique method got t implemented.

Lu, et al., [5] proposed the similarity between the sentence based on the distance. Either long sentence or short sentence, sentence is placed a distance. Distance parameter used to find the similarity between the sentence. Between the strings weights are calculated using probabilistic technique.

XiaoyanCai, et al., [6] proposed various sentence and words, topics, sub topics. Based on this information clustering are take place. For example, all subtopic is extracted and finding the similarities based on that information. Existing technique based on the ranking, high ranking information's are placed in one group. This technique based on the ranking, so cluster performance always depends on ranking factor. This is the main drawback of existing technique, this problem overcome on the proposed system, here they used reinforcement approach. This proposed technique clustering and ranking are joining to gather so that either ranking or grouping updating take place any one of that criteria get satisfied. Experiments verified that propo-

sed technique bring better result than existing techniques.

Richard Khoury, et al., {7] proposed Clusters works based on the similarities between the data points based on the similarities grouping are take place. This application used in many places image, text, audio, video clustering are done based on the content or based on some similarities metrics. This paper brings the new idea on text based clustering based on part-of speech technique. Proposed technique automatically groups the information based on the above method. Same procedure may also extended based on the question pattern, based on the response either positive or negative information's are grouped. Positive responses are in one group and negative responses are another group.

## III. EXISTING SYSTEM

Existing system used fuzzy based technique this technique ae well used in pattern and image type of input files. For this type of data sets are worked based on the relationship between the given data points. Every analyze system use to find the similarities between the data points, this method done with help of matrix based technique. Items are first constructed a matrix format then each matrix value compared with other matrix value ie pair value comparison are done. This give complex computation and gives extra burden to the user community. Other technique falls on construct chart based methodology this also give burden to the user need to construct graph than information are analyzed. Both technique gives burden to the user community. This leads additional burden before data points get analyzed. Due to this time and complexity of the work get increased. Due to that efficiency of the process get reduced. Other technique like page ranking time complexity get reduced, but user never get the expected output. All this point brings the attention to the research community there is best alternative technique required for grouping the information. The major disadvantage of this process is time and space complexity because of fuzzy clustering mechanism. The existing system uses page ranking mechanism to avoid time complexity but the result is not efficient.

### 3.1 Disadvantages of Existing System

1. Users are get extra burden for constructing matrix and graph for analyze the data sets.
2. Due to this time get increased.
3. Results suffer with unsteadiness.
4. Data objects are get repeated due to this performance get degraded.
5. Techniques works well in limited data sets.

### 3.2 Experimental Design

The proposed system implements the fuzzy logic with the clustering algorithm like the existing algorithm but instead of using the page ranking mechanism the proposed system uses the N-GRAM preprocessing mechanism which includes stop word removal, stemming etc. to avoid the space and time complexity.

#### 3.2.1 Advantage of Proposed System

- Relational clustering is achieved
- Accurate Search result
- Increase in search speed
- Identifying overlapping clusters.

## IV. EXPRIMENTAL SETUP

### 4.1 Pre-Processing

Any data mining technique started with preprocessing steps, this helps the user to bring the correct data set. Based on the input can expect quality outputs. Here preprocessing steps finds the unwanted information such as comma, prepositions are not relevant to the text extraction. It is necessary before starts text comparison this information is removed from the input data sets. After removing this unwanted information integrated data sets feed into statistical technique using vector based model.

### 4.2 Discovery of similarity groups

This work finds the groups between the data points. Data points are compared based on the similarities items are grouped. Items which having higher similarities are grouped in one, next higher similarity in other group like based on the similarities items are grouped.

### 4.3 Clustering techniques

Based on the data sets different clustering technique are used to groping the information. Finding closeness between the data points k –means and K- Medoids techniques are most effective. This algorithm find the similarities between the data points and make in one group using splitting or merging concept.

### 4.4 Reduce Noise

Output performance get improved if our data sets don't have any repeated data values. Because this unwanted data's reduce the performance of any clustering out put. It is necessary user to ensure before process get starts , data sets are cleaned and not having any unwanted data's. This process mostly done initially before the process get started. Any clustering process get repeated again and again until no single unwanted data points may available.

## V. EXPERIMENTALOUTCOMES

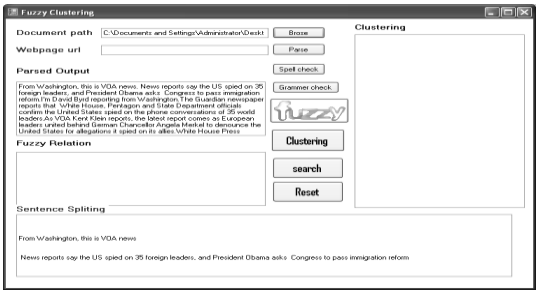


Fig 1. Pre-Processing Step

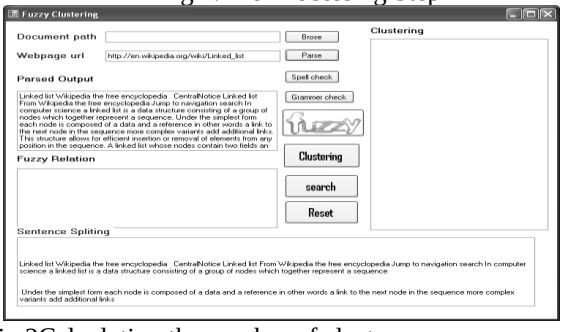


Fig 2Calculating the number of clusters

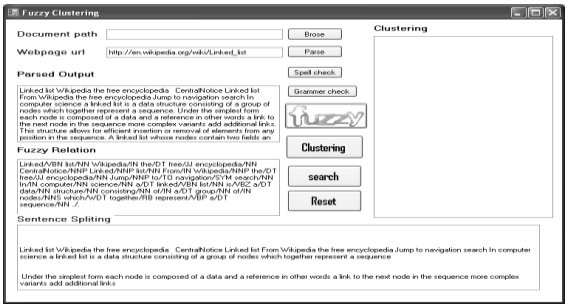


Fig 3. Clustering techniques.

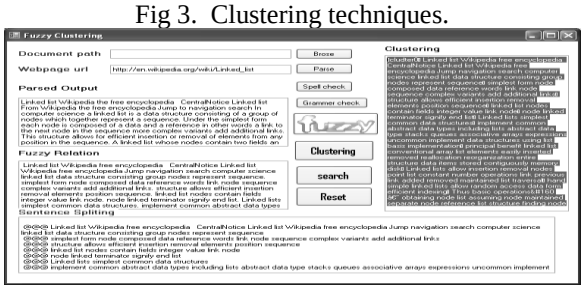


Fig 4. Cluster Formation.


Fig 5. Removing Outliers

## 6. CONCLUSION

Clustering operation depends on the input data size and quality, if our datas set consisting of pure and quality data cluster formation done effectively. Clustering can have applied to various applications it helps groups the information based on the similarities. Experimental results verified that propsed technique achieved greater performance forming number of clusters for the given input information.

## REFRENCES

[1] D. Saravanan, S. Srinivasan, Video Image Retrieval Using Data Mining Techniques. Journal of Computer Applications 5(1): 39-42 (2012).

[2] D. Saravanan, S. Srinivasan, Matrix Based Indexing Technique for Video Data, International journal of Computer Science 9(5): 534-542 (2013).

[3] Witold pedrycz, James waletzky, Fuzzy clustering with partial supervision. IEEE Transaction on Systems, Man, and Cybernetics- Part B: Cybernetics 27(5): 1- 9 (1997).

[4] Yuhu li, David Mclean, Zuhair A. Bandar, James D. O'Shea and Keely Corkett, Sentence Similiarity Based on Semantic Nets and Corpus statics. IEEE transaction on knowledge and data engineering 18(8): 1138-1150 (2006).

[5] Shin-Yee Lu, King sun Fu, A Sentence-to-Sentence Clustering procedure for pattern analysis. IEEE Systems, Man, and Cybernetics Society 8(5): 381-389 (1976).

[6] Xiaoyan Cai, Wenjie Li, You Ouyang, Hong Yan, Simultaneous ranking and clustering of sentences: A Rein forcement approach to multi-document Summarization, Proc. Of the 23rd International conference on Computational linguistics Pp. 134-142 (2010).

[7] Richard Khoury, Sentence clustering using parts-of-speech. International journal of Engineering and Electronic Business 1: 1-9 (2012).