Using Fuzzy Logic Clustering Discover Semantic Similarity in Web Document

Miss. Sukshma M. Durge
Student: Dept. of Comp. Sci. & Engg.
SGBAU, Amravati
Maharashtra, India

Mr. Yogeshwar M.Kurwade
Prof. : Dept. of Comp. Sci. & Engg.
SGBAU, Amravati
Maharashtra, India

Dr. Vilas M. Thakare
HOD: Dept. of Comp. Sci. & Engg.
SGBAU, Amravati
Maharashtra, India

Abstract— *The complex and high interactions between terms in documents demonstrates vague and ambiguous meanings. There exist complicated associations within one web document and linking to the others. Most of these approaches perform similarity and feature section methods. There is need of complex document clustering and produced meaningful document. This paper proposed methodology is capable of handles more naturally situations and improves performance.*

Keywords— Document parsing, Key term extraction, Clustering, Document Filtering, Similarity measure.

I. INTRODUCTION

The users search engines such as Google return a long ranked list of search results based on the relevance to the query terms and users have to sift through the returned list and select the desired information by browsing the title and snippet of every document. This process may be good for certain search tasks but less effective and efficient for ambiguous queries for the reason that the results on different subtopics or meanings of a query will be mixed together in the list. To address the above challenges, some effective approaches used or proposed.

Fuzzy logic models, called fuzzy inference systems, consist of a number of conditional "if-then" rules. The designer who is understands the system, these rules are easy to write, and necessary for many rules can be supplied to describe the system adequately (although typically only a moderate number of rules are needed). Fuzzy logic can handle incomplete data and problems with imprecise and it can functions model nonlinear of arbitrary complexity [1][3]. The clustering separates entities that may provoke burst from those who have less impact on the web. The remain generic to use transfer learning techniques for both entity clustering and document classification [2]. A two-class SVM was trained using those features extracted for synonymous and non-synonymous word pairs selected from WordNet synsets [4]. The hierarchical relationship organized into groups according to similarity documents measure and similarity between term and document vectors [5].

In this paper, numbers of different clustering methods, including suffix array clustering, sequential clustering algorithm, Agglomerative clustering, Most of these approaches perform similarity and feature section methods.

All these methods are compared with their quality and computational time and also accuracy. Most of these approaches perform web document clustering, is concerned with grouping objects together that are similar to each other and dissimilar to the objects belonging to other clusters The clusters produced by the clustering procedure called “hard” or “crisp” clusters, since any feature vector x either is or is not a member of a particular cluster. This is contrast to “soft” or “fuzzy” clusters, in which a feature vector x can have a degree of membership in each cluster. The proposed method extends the idea of AFCC clustering method. We can readily apply fuzzy logic k-Means algorithms. Since weighting features and dimension reduction can readily be calculated in the web document clustering. This paper tries to overcome the problem of previous methods using proposed methodology.

II. BACKGROUND

In this paper, clustering Abstract Fuzzy Combination of Criteria (AFCC) [1], Entity web popularity [2], fuzzy linguistic topological space along with a fuzzy clustering algorithm.[3], Empirical method to estimate semantic similarity [4] and Suffix Array Similarity Clustering (SASC) [5].

The Abstract Fuzzy Combination of Criteria (AFCC)[1]. Of the datasets features are distributed differently. This document representation process split into three stages the first stage selection of feature sources, then weighing features each feature and dimensionality reduction. And the limitations of FCC, EFCC and AddFCC to deal with varying term distributions across different datasets now

delve into alternative considerations that further exploit these characteristics which ultimately leads to the definition of AFCC.

The approach uses one model per cluster of entities based on the entity web popularity and also introduces different strategies for automatic classification model selection and test on the Knowledge Base Acceleration (KBA) framework from TREC [2].

The proposed fuzzy linguistic topological space along with a fuzzy clustering algorithm to discover the contextual meaning in the web documents.[3]. There is semantic tree represents the CONCEPT hierarchy and the root is the user query. Except the root, all terms on the other nodes can be considered to be advanced search queries. Then a query of a node is aggregated with other co-occurring entities to become more primitive on its branches of the tree.

This proposed empirical method is measure semantic similarity using page counts and text snippets retrieved from a web search engine for two words.[4] A considers both page counts and lexical syntactic patterns extracted from snippets. A lexical pattern extraction algorithm to extract numerous semantic relations exists between two words and a sequential pattern clustering algorithm to identify different lexical patterns that describe the same semantic relation and also both page counts-based co-occurrence measures and lexical pattern clusters used to define features for a word pair.

The Suffix Array Similarity Clustering (SASC) for clustering web search results.[5], SASC method get the document list from the search result and each document are parsed and split into sentence according to punctuations. Then creates clusters by adopting improved suffix array, which ignores the redundant suffixes, and computing document similarity based on the title and short document snippet and organized into groups according to the hierarchical relationship.

The rest of this paper is organized as follows. In Section III, we discuss previous work, Section IV, we discuss existing methodologies, Section V, analysis and discussion. Section VI we present the proposed methodology, and we present possible outcome and results in Section VII. We conclude this paper in Section VIII

.

III. PREVIOUS WORK DONE

Over the past decades, many clustering algorithms have been proposed, including suffix array clustering, sequential clustering algorithm, Agglomerative clustering, Most of these approaches perform similarity and feature section methods.

Garcia-Plaza et al. (2016) [1] has worked on fuzzy term weighing approach that makes the most of the HTML structure for document clustering called Abstract Fuzzy Combination of Criteria (AFCC) adapt datasets features are distributed differently. This solution could return not only the flexible of cluster produced but also achieve good performance.

Bouvier et al. (2015) [2] has worked on uses one model per cluster of entities based on the entity web popularity and also introduce different strategies for automatic classification model selection and test on the Knowledge Base Acceleration (KBA) framework from TREC.

Chiang et al. (2015) [3] has worked on fuzzy linguistic topological space based on associations and proposed fuzzy clustering algorithm to discover the contextual meaning in the web documents. This solution produced a significant improvement when the numbers of selected clusters become higher.

Bollegala et al. (2011) [4] has worked on empirical method to estimate semantic similarity using page counts and text snippets retrieved from a web search engine for two words. This solution could return community mining task show the ability of the proposed method to measure the semantic similarity between not only words but also between named entities, for which manually created lexical ontologies do not exist or incomplete.

Bai et al. (2010) [5] has worked on Suffix Array Similarity Clustering (SASC) for clustering web search results.

IV. EXISTING METHODOLOGIES

A. ABSTRACT FUZZY COMBINATIONS OF CRITERIA

The membership degree of object to a particular class is defined by human experts and a membership function. The term frequency in the document is defined in three fuzzy sets: low, medium, and high. The term frequency within the <title> tags. It is defined in two fuzzy sets: low and high. The emphasized is a parts of the text in the term frequency. It is composed of three fuzzy sets: low, medium and high. It is obtained by means of an Auxiliary Fuzzy System that takes as input all the positions of a term within a document (captured by the other linguistic variable term position) and returns the global position value in terms of two fuzzy sets, standard and preferential. The output of the fuzzy system and equates to the estimated importance of a term in the document content. It has five homogeneously distributed fuzzy sets: no, low, medium, high and very high.

B. ENTITY POPULARITY BASED ON TIME FEATURES

A correlation coefficient ρF,C computed using each feature vector F ∈ [F1, ..., Fn] and the class vector C. The assumption is that entities sharing the same correlation between a time feature and the document class are more likely to be equally popular. Thus deduce the popularity of an entity with other entities. Then define the correlation matrix P. There is no correlation between the categories popular/non-popular and the number of entities. A number of clusters that increases according to the number of entities is counterintuitive. The n represent the number of entities.

$$k \approx \sqrt{n/2}$$

The meta features defined use an Entity Profile EP to keep track of information related to an entity and compute similarity and divergence features between a document and different structures within the EP.

C. *Fuzzy latent semantic clustering*

The latent semantics discover a text corpus from a fuzzy linguistic perspective. Fuzzy linguistic coefficient is given to measure the possibilities of a term belonging to every category where the term is associated with other co-occurring terms. The features in a document are extracted by using semi-supervised learning schemes called named entities. The NER identify one item from a set of features that have similar attributes i.e. named categories. The Hierarchical fuzzy linguistic topological space model discover the hierarchy of semantics from a collection of documents. The central notion is n-simplex called semantics. It is based on a fuzzy characterization of the coincidence concept and obtained means of several conjunction functions for handling linguistic weighted information from every tuple of elements in the n-simplex. The fuzzy latent relational measure of every element in the simplex to be 1 that means they are coincident with their corresponding named categories after two simplexes have been merged.

D. *Semantic Similarity Measure*

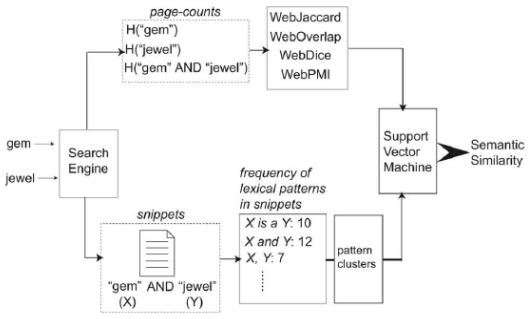

Fig. 2. Outline of the proposed method.

The Page counts-based similarity scores consider the global co-occurrences of two words on the web. They do not consider the local contexts into two words co-occur. This can be problematic if one or both words are polysemous or when page counts are unreliable. The lexical pattern extraction algorithm typically extracts a large number of lexical patterns. Clustering algorithms based on pairwise comparisons among all patterns are prohibitively time consuming when the patterns are numerous. So that a sequential clustering algorithm to efficiently cluster the extracted patterns.

$$\mu(a) = \sum_i f(P_i, Q_i, a).$$

A weight wij to a pattern ai that is in a cluster cj as follows:

$$w_{ij} = \frac{\mu(a_i)}{\sum_{t \in c_j} \mu(t)}.$$

The value of the jth feature in the feature vector for a word pair (P,Q) as follows:

$$\sum_{a_i \in c_j} w_{ij} f(P, Q, a_i).$$

E. *Suffix Array Similarity Clustering (SASC)*

Get the document list from the search result and each document are parsed and split into sentence according to punctuations. All punctuations replace by unique token and removed all HTML tags. Here they used the stemming algorithm to stem word then get the title and small snippet. The key terms extract by using Suffix Array for better performance and more meaningful phrases. The extracting terms organizing into clusters and extracting cluster labels. After finishing all the process get all final clusters.

For each cluster C_i in C
  For each cluster C_j in $C - C_i$
    If $C_i \subset C_j$ then C_i is a child of C_j
    End if
  End for
End for

Figure 3. The analytic hierarchy process.

V. ANALYSIS AND DISCUSSION

To improved accuracy and quality of web documents clustering result, it use a baseline to compute the weight of each word occurring in a document by using the well-known TF-IDF term weighing function. The unsupervised reduction method called Most Frequent Terms (MFT) is used to reduce the computational cost. The overall F1 score is computed as the weighted average of the F1 scores for each category. The qualitative analysis of the membership functions are utilizing and then test AFCC, evaluating and analyzing its performance in comparison with the techniques studied previously. [1].

The entity cluster is evaluated based on the new data found. The entity moves from its current cluster to the estimated cluster. The maximum average f -measure obtained to the official score for a system and cutoff is equal to 0.5, only scores above the cutoff are considered when computing the average f-measure. All positive documents having a score below are considered false negative and all negatives document below the cutoff are considered true negative. [2].

To implement effective clustering result, it uses Normalized Mutual Information (NMI) to calculate the normalized mutual information of the topic and its corresponding cluster accordingly. Thus tests, the documents with multiple topics (category labels) and with single topic were separated, and the topics with less than five documents were removed. This paper F1 measure used was obtained when β is set to be 1, which means precision and recall are equally weighted for evaluating the performance of clustering and the non-overlapping scenario each document belongs to exactly one cluster.[3].

Pearson and Spearman correlation coefficients to measure the semantic similarity cluster accordingly. Thus tests method achieves the highest Pearson and Spearman coefficients and outperforms all other web-based semantic similarity measures.[4].

The suffix array Similarity Clustering the key terms discovered taken as candidates for cluster labels. Then applied the various clustering algorithms to the document collections (the top 100 search results' titles and snippets) and compared their precision. Then measure similarity between term and document vectors. The highest total score will be the only label of the cluster.[5].

A brief comparison of AFCC Clustering, Entity web popularity, FLSC, SSM and SASC are as shown in table.

Table 1: COMPARISON BETWEEN ASKM, SSC, TW-K-means, TRIPARTITE, AND CONSTRAINED CLUSTERING

Clustering methods	Advantages	Disadvantages
AFCC Clustering	1. Increase in cluster flexibility. 2. Adapt the representation to different datasets that could have different features.	1. The clustering quality reduced.
Entity web popularity	1.Clustering more efficient. 2.Scalable.	1. Cluster to a group entity.
FLSC	1.Increase in cluster purity 2.The document contains multi-topics and more heterogeneous.	1. single term is not able to identify a latent concept in a document.
SSM	1. Increase in cluster accuracy. 2. Improving mining task. 3. Achieving high correlation with human ratings.	1. Not adequately covered by manually created resources.
SASC	1. Better performance. 2. Better cluster description quality. 3. Lower computational requirement by reducing the times of matrix operations. 4. Higher accuracy.	1. Time consuming .

VI. PROPOSED METHODOLOGY

By applying fuzzy logic k-means clustering algorithm to address the existing problems.

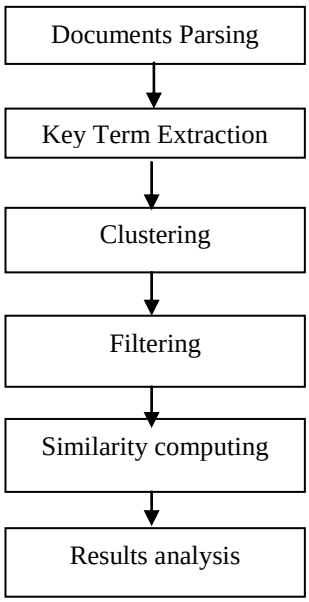

Fig.1. Flowchart of the proposed algorithm

VII. POSSIBLE OUTCOMES AND RESULTS

The fuzzy logic K-Means clustering produce k number of clusters and compute the similarity between words to gives the efficient hierarchical clusters. The performance of clustering result is improved and less time consuming.

VIII. CONCLUSION

This paper exhausted different clustering algorithm for including suffix array clustering, sequential clustering algorithm, Agglomerative clustering, Most of these approaches perform similarity and feature selection methods. The results of these clustering algorithms the proposed algorithm show that the algorithm is able to achieve superior performance. The similarity between objects is based on a measure of the distance between them. The clusters are semantically related words.

Acknowledgment

I, Sukshma M. Durge, thankful of my Prof. Y. M. Kurwade, and Supervisor Dr. V. M. Thakare for helping to complete this research paper.

References

[1] A. P. Garcia-Plaza, V. Fresno, R. Martinez and A. Zubiaga, "Using Fuzzy Logic to Leverage HTML Markup for Web Page Representation", IEEE TRANSACTIONS ON FUZZY SYSTEMS, pp. 1-17, JUNE 2016.

[2] Vincent Bouvier and Patrice Bellot, "Use of Web Popularity on Entity Centric Document Filtering", 2015 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 268-275, 2015.

[3] I-Jen Chiang, Charles Chih-Ho Liu, Yi-Hsin Tsai, and Ajit Kumar, " Discovering Latent Semantics in Web Documents Using Fuzzy Clustering", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL.23, NO. 6. pp. 2122-2134, DEC. 2015.

[4.] D. Bollegala, Y. Matsuo, and Mitsuru Ishizuka, "A Web Search Engine-Based Approach to Measure Semantic Similarity between Words", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 7, pp. 977-990, JULY 2011.

[5] Shunlai Bai, Wenhao Zhu, Bofeng Zhang, Jianhua Ma, "Search Results Clustering Based on Suffix Array and VSM", IEEE/ACM International Conference on Green Computing and Communications & on Cyber, Physical and Social Computing, pp. 852-857, 2010.