www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume – 5 Issue -03 March, 2016 Page No. 15991-15995

Web Search Result Clustering using IFCWR Algorithm

Arun Kumar Pal[1], Ashwini Londhe[2], Nitin Patil[3], Vinod Bankar[4]

[1]Student (UG), Department of Computer Engineering,
A.I.S.S.M.S College of Engineering, Pune, Maharashtra, India
arunpal2626@gmail.com

[2]Student (UG), Department of Computer Engineering,
A.I.S.S.M.S College of Engineering, Pune, Maharashtra, India
Ashwinilondhe1111@gmail.com

[3]Student (UG), Department of Computer Engineering,
A.I.S.S.M.S College of Engineering, Pune, Maharashtra, India
Nit.rinku@gmail.com

[4]Student (UG), Department of Computer Engineering,
A.I.S.S.M.S College of Engineering, Pune, Maharashtra, India
bvinod60@gmail.com

Abstract: ***In the modern age of Internet, usage of social media is growing rapidly on internet, organizing the data, interpreting and supervising User generated content (UGC) has become one of the major concerns. Updating new topics on internet is not a big task but searching topics on the web from a vast volume of UGC is one of the major challenges in the society. In this paper we deal with web search result clustering for improving the search result returned by the search engines. However there are several algorithms that already exist such as Lingo, K-means etc. In this paper basically we work on descriptive-centric algorithm for web search result clustering called IFCWR algorithm. Maximum numbers of clusters are randomly selected by using Forgy's strategy, and it iteratively merges clusters until most relevant results are obtained. Every merge operation executes Fuzzy C-means algorithm for web search result clustering. In Fuzzy C-means, clusters are merged based on cosine similarity and create a new solution (current solution) with this new configuration of centroids. In this paper we investigate the Fuzzy C-means algorithm, performing pre-processing of search query algorithm and try to giving the best solution.***

Keywords: Iterative Fuzzy C-means, IFCWR, Query pre-processing algorithm, Porters Stemming algorithm.

1. Introduction

In the present scenario, web search result or web document clustering has become a very interesting and the challenging task in computer researches involving web searches and information retrieving from web. Clustering Web document is an approach to increase amount of relevant documents presented for user to review, while reducing time spent reviewing them. Such systems are called Web clustering engines and some already existing systems are Carrot2, SnaketT, Yippy (initially named Vivisimo then Clusty), iBoogie and KeySRC [1].

In this paper we study the methods related with clustering-pattern like similarity cascades. To organize the data into a meaningful and effective manner the technique is use is called TDT (Topic detection and tracking). The task of web topic detection is discovering of a tiny fraction of web pages firmly connected by a crucial event from a large amount of social media.

For generating effective result from clustering of web documents, the algorithm must fulfill the following basic needs [1,4]: Auto-generating number of clusters to be created, create relevant clusters for the user and assign appropriate clusters to each document; define labels -or names- for the clusters that are easy to understand; handle clusters overlapping (i.e. documents can belong to more than one clusters); handle the time to process (the algorithm must not just work with full text of web document but also with the snippets); and handle the noise that are most common in the document collection. Even though there are several algorithms that already exist and are implemented but still according to researches there is place for more efficient algorithms. Algorithms are classified into three types [2]: description-aware, description-centric and data-centric. They all build clusters of documents and most of them assign each group a label.

The paper consists of following sections. Section 2 describes some earlier related models and work. Section 3 describes about the IFCWR algorithm and its major inclusion steps. Section 4 describes the proposed system's block diagram and its results.

2. Literature Survey and Related Work

2.1 Data-centric algorithm

Data-centric algorithm is basically used for partitioned, hierarchical, fuzzy data clustering. They search the best solution for a data clustering, but they have failed to find strong labeling view or in the explanation of obtained groups (clusters). In partitioned clustering, the mostly recognized algorithms are: k-means, k-medoids, and Expectation Maximization. In fuzzy clustering, a new method using FTCA (fuzzy transduction-based clustering algorithm) was presented in 2010 [5]. FTCA results are impressing but are

not measured over known datasets which is necessary to effectively check the algorithm's results.

2.2 Description-aware algorithm

Description-aware algorithms emphasizes to one specific feature of the clustering than to the rest. For example, giving priority to labeling of groups and achieve results that are easily understood by user. This however, decreases their quality in the process of cluster creation. Suffix Tree Clustering (STC) [4] is an example of this type of algorithm, which continuously creates labels easy to understand by users, based on common phrases appearing in the documents.

2.3 Description-centric algorithm

Description-centric algorithms [2, 6-11] are devised specifically for clustering of web document, searching a balance between the cluster quality and their description (labeling). For example Lingo [7] (realized by Carrot2 in 2001), which makes use of Singular Value Decomposition (SVD) to find the best homogeneity between terms, but groups the documents on the basis of most frequent phrases in the document collection.

2.4 Lingo algorithm

With massive growth of the internet it has become laborious to review the relevant document in response of the users query. Currently available search engine return the ranked list along with the snippets [12]. The Carrot2 is the recent project which has implemented Lingo to cluster web search results. They fetch the data from the Google servers and arrange all the links together for increasing the speed of the search. The idea behind the lingo is that first identify meaningful labels for clusters and then as per the labels identify the actual content of the groups. In Lingo, Vector space model (VSM) is used for representing the multidimensional vector. Every component of the vector represents a particular keyword or term connected with the given document. The value of every component depends on the degree of similarity between its related term and the respective document. Several methods for calculating the relationship- also referred as weight of term- have been proposed.

3. Methodology and algorithm description

Clustering of Google search results is achieved using IFCWR algorithm which depends on the results by the search engine. However, to get the most relevant results we need to provide a well processed input string. To achieve that, we apply various algorithms on the input query. This is called preprocessing of search query. Following are the preprocessing steps/algorithms used:

3.1 NLP- Natural Language processing

Natural language processing involves finding main targeted answers to a query. For example, the query *'Which city has highest literacy rate?'*, is handled by a standard search engine based on the keywords *'city', 'highest', 'literacy'* and *'rate'*. A natural language search engine would rather try to understand the nature of the question and then search to get a subset of web containing the answer [13].

3.2 Lower Case Filtering

Lower case filtering allows us to ignore the case sensitive results to be treated as different words. E.g. 'project', 'PRoject' and 'Project' are treated as similar words.

3.3 Stop word removal

Stop word removal involves elimination of words like, 'the', 'is', 'or', 'and', 'a' etc. Some tools avoid this function to support searching of Phrase.

3.4 Porter's Stemming algorithm

The Porter Stemming algorithm (or 'Porter stemmer') is a process for eliminating the commoner morphological and in flexional endings from words in English [14]. Here a words are reduced to stem words (refer Table 1).

Table 1: Porter Stemming of words.

Word	Stem word
abandon	Abandon
Abandoned	Abandon
Abate	Abat
Abated	Abat
Abatement	Abat
Abatements	Abat
Abates	abat

3.5 Applying IFCWR algorithm

(1). **Initialize algorithm parameters.** The algorithm is initialized with the parameters Maximum Time for Execution (MTE) or the Maximum Iterations (MI). These parameters control the execution of the iterative process in the algorithm. For web results clustering usually a MTE value is 2 seconds.

(2). In **Document preprocessing** a TDM view of document is used. In IR, TDM is globally used document representation structure. It is depending on the method called vector space model [1, 3, 4, 15]. In this model, the documents are arranged as bags of words, the collection of document is expressed by a matrix of D-terms by N-documents, each document is expressed by a vector of normalized term frequency ($freq_i$) by the inverse frequency of document for that term (represented by equation (1)), and the degree of likeliness between documents, or between a document and the cluster-center or between a document and the user's search query is measured by the cosine similarity formula.

$$w_{i,j} = \frac{freq_{\cdot i,j}}{\max(freq_{\cdot i,j})} \times \log\left(\frac{T}{t_j}\right) \qquad (1)$$

Where $freq_{\cdot i,j}$ is the calculated frequency of term j in document i, $\max(freq_{\cdot i,j})$ is the max. identified frequency in the document i, T is the total number of web-documents in the collection, and t_j is the number of documents where term j is present.

(3). When the algorithm creates the **Initial Solution**, it selects an initial number of clusters based on the number of web-documents. This values is equivalent to $\lceil\sqrt{T}\rceil$, where T

represents the number of web-documents, which cannot be less than eight (8) nor greater than T.

(4). Select the **best possible solution**: Find and select the best possible solution from the best solution's list. The prime solution is the solution with the lowest fitness value (minimize Bayesian Information Criteria, see equation (2)). Return this solution as the best clustering solution.

$$BIC = k \times Ln(n) + n \times Ln\left(\frac{E}{n}\right) \quad (2)$$

$$E = \sum_{j=1}^{k} \sum_{i=1}^{n} P_{i,j} \left\| x_i - c_j \right\|^2 \quad (3)$$

Where E is the Sum of Squared Error, $P_{i,j}$ is the degree of association of term x_i with cluster j, x_i is the i-th term of m-dimensional measured data, c_j is the m-dimensional center of the cluster, and ||*|| be any norm expressing the similarity between any measured data and the center.

(5). Assigning **labels** to the clusters: Using a Frequent PHrases (FPH) approach for labeling every cluster. In IFCWR labeling of clustering is done by following various steps like Conversion of representation scheme(from character-based representation to word-based), Document concatenation, Complete phrase discovery (Right-complete phrases and left-complete phrases), Final selection is based on the Threshold Term Frequency, Creation of the "Others" label if documents do not succeed to meet the Threshold Term Frequency.

4. Proposed System

Clustering of search results is implemented here using technologies like Jsp, MySQL, AJAX etc. In this system the user interacts using a browser to perform query. A set of N documents is fetched by Google server after preprocessing of the search string. The system performs IFCWR algorithm until the best result is obtained. A well structured result set, that is, quality of cluster as well as description of each group is returned to the user.

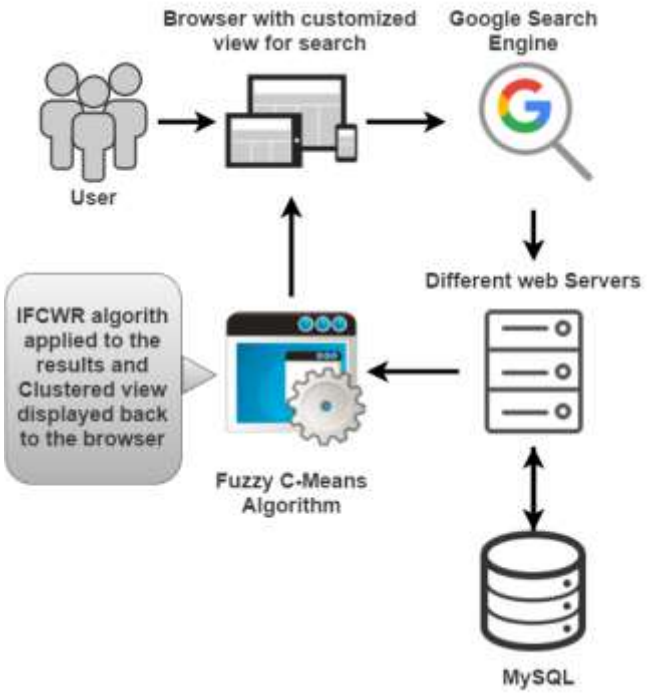

Figure 1: Schematic Diagram of Implemented System and its working.

Fuzzy c-means (FCM) is a clustering method that allows some data to belong to more than one cluster.

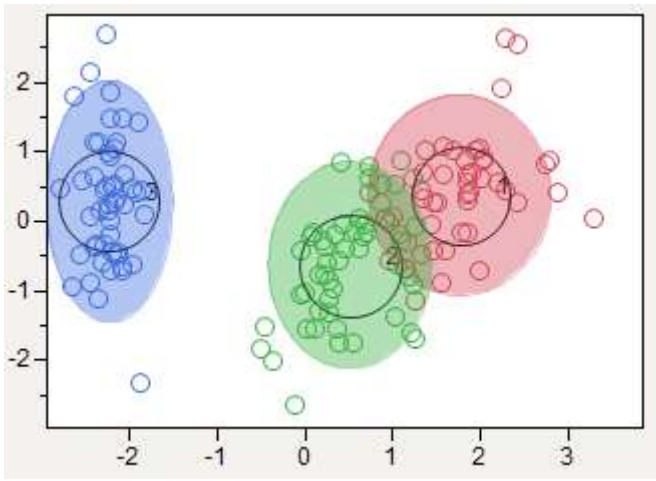

Figure 2: The overlapping regions shows data belongs to more than one cluster.

This algorithm works by allocating associate-ship to every data point corresponding to every cluster center on the basis of the cluster center and the data point distance. The closer the data is to the cluster center more is its membership towards that particular center. Clearly, addition of membership of every data point should be equal to one. After every iteration, cluster centers and the association of point with them are updated as per the Equation (3).

DOI: 10.18535/ijecs/v5i3.19

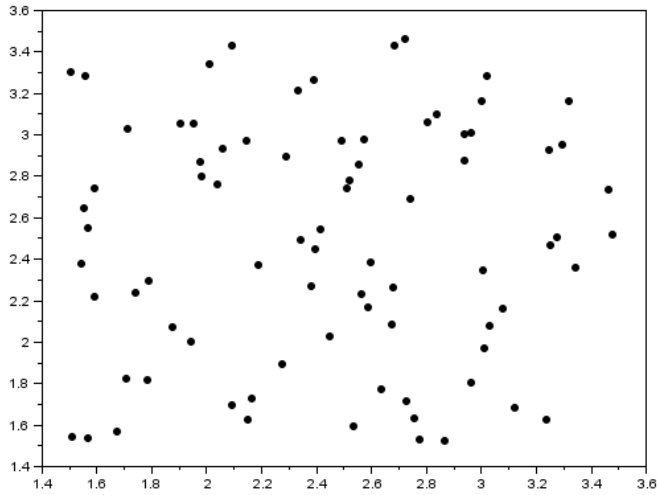

Figure 3: Data set before applying fuzzy C-means algorithm.

The algorithm is composed of the following steps:

Step 1. Allocate initial values to $U=[u_{ij}]$ matrix, $U^{(0)}$

Step 2. At k-iteration: determine the center vector $C^{(k)}=[c_j]$ with respect to $U^{(k)}$

$$C_j = \frac{\sum_{i=1}^{N} u_{i,j}^{m} \times x_i}{\sum_{i=1}^{N} u_{i,j}^{m}} \qquad (4)$$

Step 3. Update vectors $U^{(k)}$ and $U^{(k+1)}$ using,

$$u_{i,j} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \qquad (5)$$

Step 4. If $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ then terminate the loop; otherwise go to *step 2.*

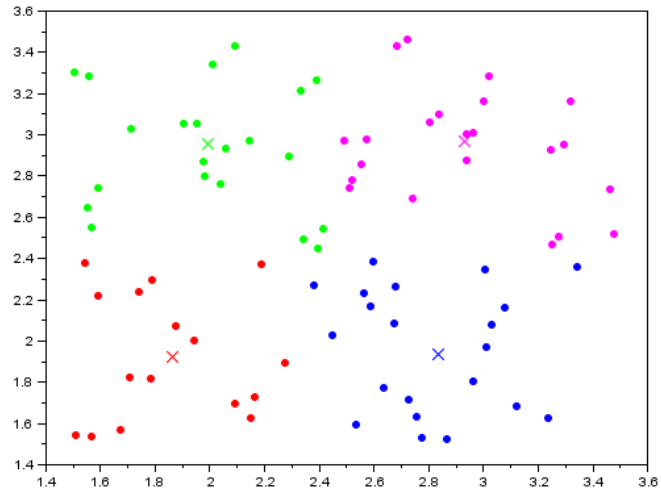

Figure 4: Data set after applying fuzzy C-means algorithm. (× refers to the cluster center)

We apply fuzzy c-means clustering algorithm on collected objects for making the customized window on browsers for user search quickly. Figure shows a clustered view of applied algorithm.

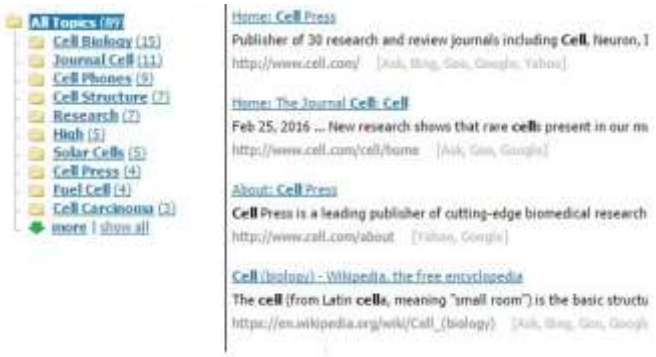

Figure 5: Clustered view with keyword "cell".

In comparison with Lingo and STC, the number of clusters calculated is much better in IFCWR . IFCWR finds on average less number of sub-topics as compared to STC and Lingo on both AMBIENT and MORESQUE dataset [16].

5. Conclusion

IFCWR algorithm has been successfully designed and implemented. It is a description centric algorithm based on Fuzzy C-means to cluster the web search results and generate N number of clusters, where N is automatically defined. In comparison with Lingo and STC, the number of clusters calculated is much better in IFCWR . IFCWR finds on average less number of sub-topics as compared to STC and Lingo on both AMBIENT and MORESQUE dataset. As a part of data mining, clustering of web searches using IFCWR algorithm yields fewer labels which are more relevant and based on common phrases. However there is scope for future development to increase the efficiency of the algorithm by machine learning approach and knowledge storage to create labels effectively and make searching less time consuming.

.

References

[1] C. Carpineto, et al., "Evaluating subtopic retrieval methods: Clustering versus diversification of search results," Information Processing & Management, vol. 48, pp. 358-373, 2012.

[2] C. Carpineto, et al., "A survey of Web clustering engines," ACM Comput. Surv., vol. 41, pp. 1-38, 2009.

[3] R. Baeza-Yates, A. and B. Ribeiro-Neto, Modern Information Retrieval: Addison-Wesley Longman Publishing Co., Inc., 1999.

[4] Z. Oren and E. Oren, "Web document clustering: a feasibility demonstration," presented at the Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, 1998.

[5] T. Matsumoto and E. Hung, "Fuzzy clustering and relevance ranking of web search results with differentiating cluster label generation," in Fuzzy Systems (FUZZ), 2010 IEEE International Conference on, 2010, pp. 1-8.

[6] T. Matsumoto and E. Hung, "Fuzzy clustering and relevance ranking of web search results with differentiating cluster label generation," in Fuzzy

Systems (FUZZ), 2010 IEEE International Conference on, 2010, pp. 1-8.
[7] S. Osiński and D. Weiss, "A concept-driven algorithm for clustering search results," Intelligent Systems, IEEE, vol. 20, pp. 48-54, 2005.
[8] D. Zhang and Y. Dong, "Semantic, Hierarchical, Online Clustering of Web Search Results," in Advanced Web Technologies and Applications, ed, 2004, pp. 69-78.
[9] B. Fung, et al., "Hierarchical document clustering using frequent itemsets," in Proceedings of the SIAM International Conference on Data Mining, 2003, pp. 59-70.
[10] G. Mecca, et al., "A new algorithm for clustering search results," Data & Knowledge Engineering, vol. 62, pp. 504-522, 2007.
[11] F. Beil, et al., "Frequent term-based text clustering," in KDD '02: International conference on Knowledge discovery and data mining (ACM SIGKDD), Edmonton, Alberta, Canada, 2002, pp. 436-442.
[12] Junbiao Pang, et al. "Unsupervised Web Topic Detection Using A Ranked Clustering-Like Pattern Across Similarity Cascades," in IEEE Transactions on Multimedia, Vol.17, No.6, JUNE 2015, pp. 843-853.
[13] Anonymous, "Natural Language Processing," [Online]. https://en.wikipedia.org/wiki/Natural_language_processing.
[14] M. Porter, "The Porter Stemming Algorithm," [Online]. Available: http://tartarus.org/martin/PorterStemmer/.
[15] K. Hammouda, "Web Mining: Clustering Web Documents A Preliminary Review," ed, 2001, pp. 1-13.
[16] C. Cobos et al., "Clustering of web search results based on an Iterative Fuzzy C-means Algorithm and Bayesian Information Criterion" in IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint, Edmonton, AB , DOI: 10.1109/IFSA-NAFIPS.2013.6608452, pp.507-512.
[17] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," Biometrics, vol. 21, pp. 768-769, 1965.

Author Profile

Arun Kumar Pal, pursuing Bachelor's degree from Savitribai Phule Pune University in A.I.S.S.M.S College of Engineering, Pune, Maharashtra, India.

Ashwini Londhe, pursuing Bachelor's degree from Savitribai Phule Pune University in A.I.S.S.M.S College of Engineering, Pune, Maharashtra, India.

Nitin Patil, pursuing Bachelor's degree from Savitribai Phule Pune University in A.I.S.S.M.S College of Engineering, Pune, Maharashtra, India.

Vinod Bankar, pursuing Bachelor's degree from Savitribai Phule Pune University in A.I.S.S.M.S College of Engineering, Pune, Maharashtra, India.