Volume 6, Issue 8, August 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Hierarchical Clustering Approach in the Retrieval of Ontologies from Semantic Web Repositories

Dr. K. Sridevi
Assistant Professor, Department of Computer Science, Nehru Memorial College,
Puthanampatti, Trichy District, Tamilnadu, India

Abstract- Semantic Web is an extension of current Web which adds a new structure to the current Web. Ontology has a major role in Semantic Web development. Since the size of Ontology repositories are getting increased, the process of retrieving more exact ontologies in efficient way becomes a tough issue. The retrieval of relevant ontologies and ranking them accordingly are required much to save searchers time. The integration of hierarchical clustering approach along with an existing partitioned clustering helps in retrieving results from the repository quicker. This paper suggests an integration of hierarchical clustering in every partitional cluster which improves the retrieval of results in the semantic web repository. This approach enhances the convenience of users with and reduced time complexity in finding the relevant needs of the searcher.

Keywords- Semantic Web, Ontology, Partitional Clustering, Hierarchical Clustering.

I. INTRODUCTION

Conventional direct keyword based information retrieval mechanism cannot meet the growing user retrieval need. The keyword based information retrieval technology fails to integrate information spread over different resources. This technology does not use the semantics, to overcome this problem in Web, the next-generation Web, which Tim Berners-Lee and others call the “Semantic Web,” aims at allowing machines to process information automatically and which also gives focus on semantics of the content [1]. Ontologies offer an efficient way to reduce the amount of information overload by encoding the structure of a specific domain and offering easier and meaningful access to the information for the users [2]. There are number of ontology search engines with which, it is possible to search for the need. The search engines also employ ranking mechanism which makes the user to get their more relevant ontology. But still there are researches to improve the time spent on the search for getting the relevant results.

Clustering is one of the main data analysis techniques and deals with the organization of a set of objects in a multidimensional space into unified groups, called clusters. Each cluster contains objects that are very similar to each other and very dissimilar to objects in other clusters. Cluster analysis aims at discovering objects that have some representative behaviour in the collection. The Information Retrieval community has explored document clustering as an alternative method of organizing retrieval results. Grouping similar documents together into clusters will help the users find relevant information quicker and will allow them to focus their search in the appropriate direction. Various clustering techniques are now being used to give meaningful search result on web [3].

The distinction among different types of clustering is whether the set of clusters is nested or unnested, or in more traditional terminology, hierarchical or partitional. A partitional clustering is simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset. If clusters are permitted to have sub clusters, then it is hierarchical clustering, which is a set of nested clusters that are organized as a tree. Each node (cluster) in the tree (except for the leaf nodes) is the union of its children (subclusters), and the root of the tree is the cluster containing all the objects. Finally, a hierarchical clustering can be viewed as a sequence of partitional clusterings and a partitional clustering can be obtained by taking any member of that sequence, by cutting the hierarchical tree at a particular level. This work focuses on the integration of hierarchical clustering with partitional clustering.

The remainder of this paper is organized as follows. The next section reviews the related works carried out on retrieving ontologies with clustering approach. Section 3 describes the methodology of ontology retrieval using hierarchical clustering. Section 4 presents the results and discussion of the recommended methodology. Section 5 explores the conclusion.

II. RELATED WORKS

The reference paper [4] proposes a rough k-means clustering algorithm which enables users to effectively mine web logs records to discover interesting user access patterns. The paper [5] focuses on web usage mining using classification and clusteirng, the key process of extracting knowledge of user access pattern from web servers. In the context of document retrieval, the hierarchical algorithm seems to perform better than the partitional algorithms for retrieving relevant documents as referenced in [6]. Similarly, Larsen [7] also observes that group average greedy agglomerative clustering outperformed various partitional clustering algorithms in document data sets from TREC and Reuters.

The experimental evaluation of Ying Zhao shows that for seven different global criterion functions, partitional algorithms always lead to better clustering results than agglomerative algorithms, which suggests that partitional clustering algorithms are well-suited for clustering large document datasets due to not only their relatively low computational requirements, but also comparable or even better clustering performance as represented in the paper [8]. The paper referenced in [9] focuses on presenting a performance analysis of various techniques available for document clustering. The study explored in reference [10] presents the processes of creating taxonomy for a set of journal articles using hierarchical clustering algorithm and complete hierarchical clustering was used to create a cluster of the articles. The work explored in [11] gives the survey and review of four major hierarchical clustering algorithms called CURE, ROCK, CHAMELEON and BIRCH. The obtained state of these algorithms helps in eliminating the current problems as well as deriving more robust and scalable algorithms for clustering.

III. METHODOLOGY

3.1 Hierarchical clustering

Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample.

Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types.

- Agglomerative: This is a "bottom up" approach. Each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive: This is a "top down" approach. All observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

***Agglomerative Hierarchical Algorithm*:**

Every document is considered as a separate cluster. Next on the basis of similarity, clusters are combined repeatedly till single cluster is formed.

The steps can be summarized as follows.

i) Consider each document as a single cluster.
ii) Calculate similarity of cluster Ai with cluster Bj then merge the two having maximum similarity.
iii) Repeat step 2 till single cluster is formed.

The linkage criteria determine the metric used for the merge strategy.

- *Ward* minimizes the sum of squared differences within all clusters. It is a variance-minimizing approach and in this sense is similar to the k-means objective function but tackled with an agglomerative hierarchical approach.
- *Maximum* or *complete linkage* minimizes the maximum distance between observations of pairs of clusters.
- *Average linkage* minimizes the average of the distances between all observations of pairs of clusters.

Agglomerative cluster has a “rich get richer” behavior that leads to uneven cluster sizes. In this regard, complete linkage is the worst strategy, and Ward gives the most regular sizes. However, the affinity (or distance used in clustering) cannot be varied with Ward, thus for non Euclidean metrics, average linkage is a good alternative.

Divisive Hierarchical Clustering:

A top-down clustering method and is less commonly used. It works in a similar way to agglomerative clustering but in the opposite direction. This method starts with a single cluster containing all objects, and then successively splits resulting clusters until only clusters of individual objects remain.

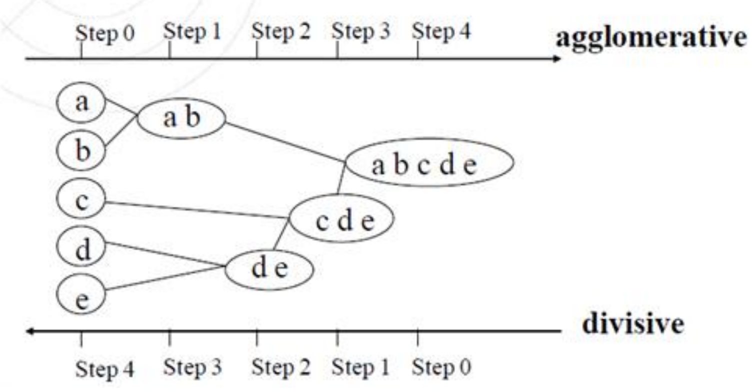

Figure 1. Two major approaches of Hierarchical Clustering

3.2 Ontology Retrieval and Ranking with Hierarchical Clustering Method

Retrieving more relevant ontologies quickly from the ontology repositories and ranking will make the searcher to get the accurate results on the top list quickly. The proposed system depends on hierarchical clustering techniques along with an existing partitional clustering to return the processed results of Swoogle search engine from the log. Then ranking is applied according to the relevancy and hierarchy of clusters retrieved. Figure 2 shows the architecture of the proposed system.

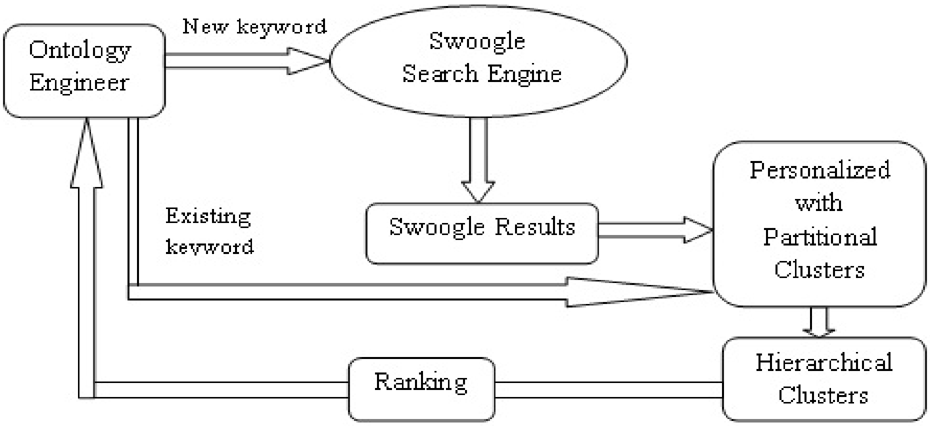

Figure 2. An Architecture of the Proposed System with Integrated Hierarchical Clustering

This system receives the keyword for searching the ontology from the ontology engineer, if the keyword is new, which will be submitted to the Swoogle search engine and when hit is made, personalization and clustering are carried out.

Partitional clusters are further clustered with hierarchical clustering to improve the performance of classification further. Finally, the revised result is returned back to the ontology engineer. This work enables the searchers an ease way of attaining their needs quickly.

Clusters are framed for similar users using usage data (UserId, Date and time of access, accessed URL, Search keyword) such as UC_{k1}, UC_{K2}, UC_{K3}, ….. $UC_{K.}$ where UC_{k1} is the User Cluster for Keyword 1, UC_{k2} is the User Cluster for Keyword 2 and so on as shown in the figure 3.

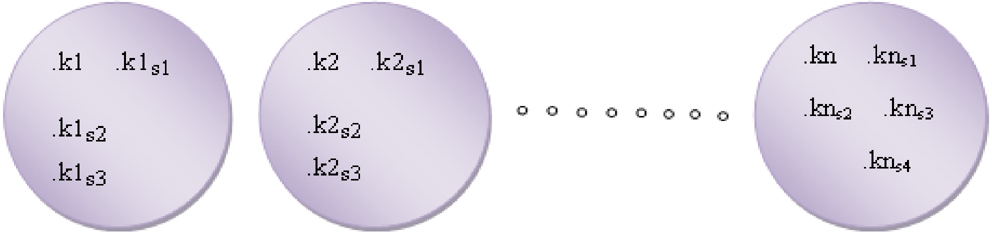

Figure 3. User Clusters (UC) based on keywords

In the figure 3, k1 represents keyword 1, k2 represents keyword 2, and so on. $k1_{s1}$ represents synonym 1 of the keyword 1, $k1_{s2}$ represents synonym 2 of the keyword 1 and so on. Similar type of keyword and its synonyms are all clustered under one user cluster which means, the users in the cluster UC_{k1} are all similar type of users.

These user clusters are further clustered to improve the performance further using hierarchical clustering. This classifies the each cluster objects using hierarchical clustering method based on the methodology represented in the section 3.1.

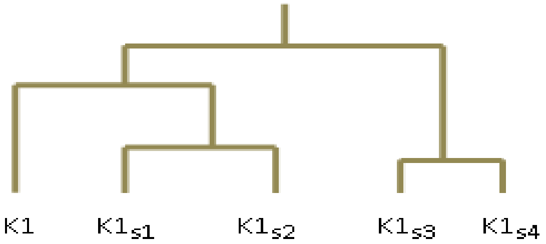

Figure 4. User Clusters (UC) based on keywords

A sample dendrogram of the cluster UC_{k1} is shown in figure 4. As per the query request, the hierarchical tree can be cut down at any level and the corresponding ontologies are retrieved. These ontologies are ranked first and the remaining relevant ontologies follow them. This places the highly relevant document in the top of the list.

IV. RESULTS AND DISCUSSION

This system is implemented in Net Beans environment using Java. The ontologies are processed using Jena API. Experiment has been done by using number of partitional clusters for various numbers of keywords. The performance metrics for analyzing clusters such as precision, recall and F-measure are used. The overall accuracy is observed using hierarchical clustering for different number of clusters. The measurements for precision, recall and F-measure are shown below.

$$Precision\ (p) = \frac{Relevant\ Retrieved\ Items}{Retrieved\ Items}$$

$$Recall\ (r) = \frac{Relevant\ Retrieved\ Items}{Relevant\ Items}$$

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The precision, recall and f-measure values are predict the best in this proposed method. Table 1 denotes the average measures computed for difference keywords in various domains.

Table 1. Precision, Recall and F-Measure for Partitional and Hierarchical Method

Method	*Precision*	*Recall*	*F-Measure*
Partitional	0.67	0.81	0.733
Hierarchical	0.92	0.74	0.82

The precision value is higher than the partitional method and the recall value is very low in proposed method. Also the F-measure value for proposed method is best when compared partitional method.

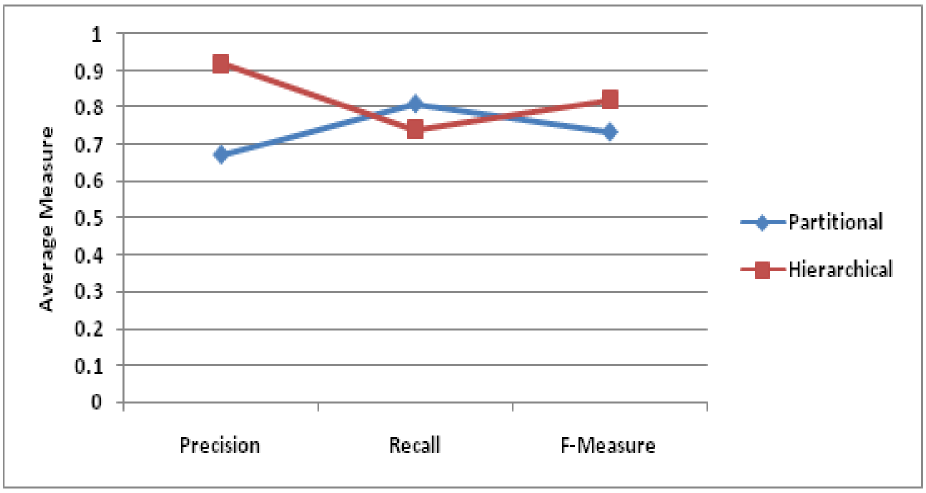

Figure 5. Graphical Representation of Performance Measures

The graphical representation as shown in figure 5 reflects the better performance results of the proposed system which integrates the hierarchical clustering approach along with an existing partitional clustering appraoch.

V. CONCLUSION

This proposed system is implemented by taking user convenience and response as primary aspect along with relevancy and time consumption while searching the content in Semantic Web. The propsosed work performs integration of hiearchical clustering algorithm with partitional clustering method and the performance is tested by applying validation measures like precision, recall and f-measure. The experimental results show that this system provides good quality of results and identifies highly relevant document quickly by following the exact hierarchy. This work can be extended by improving the relevancy of the results and reducing the time taken to process.

REFERENCES

[1] Stumme.G, Hotho.A, Berendt.B, "*Semantic WebMining: State of the art and future directions*", Web Semantics: Science, Services and Agents on the World Wide Web 4(2) 2006 124-143 Semantic Grid – The Convergence of Technologies.

[2] http://en.wikipedia.org/wiki/Ontology

[3] Oikonomakou, Nora, and Michalis Vazirgiannis. "*A Review of Web Document Clustering Approaches.*" Data Mining and Knowledge Discovery Hankbook.

[4] Peilin Shi, "*An Efficient Approach for Clustering Web Access Patterns from Web Logs*", International Journal of Advanced Science and Technology, Volume 5, April, 2009.

[5] Akshay Kansara, Swati Patel, "*Improved Approach to Predict user Future Sessions using Classification and Clustering*", International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064, Volume 2 Issue 5, May 2013.

[6] P. Willett., "*Recent trends in hierarchic document clustering: A critical review*", In Information Processing and Management, 24(5):577-597, 1988.

[7] Bjornar Larsen and Chinatsu Aone, "*Fast and effective text mining using linear-time document clustering*", In Proc. of the Fifth ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining, pages 16–22, 1999.

[8] Ying Zhao ,George Karypis, "*Comparison of Agglomerative and Partitional Document Clustering Algorithms"*, DOE ASCI program.

[9] S. C. Punitha, P. Ranjith Jeba Thangaiah and M. Punithavalli, "*Performance Analysis of Clustering using Partitioning and Hierarchical Clustering Techniques",* International Journal of Database Theory and Application Vol.7, No.6 (2014), pp.233-240.

[10] Apeh, Ayo I., Olatunde, Olabiyisi S., Owolabi, Olumide, "*A Hierarchical Clustering Approach for the Creation of a Simple Semantic Web Application*"**,** Information and Knowledge Management, ISSN 2224-5758 (Paper) ISSN 2224-896X (Online), Vol.4, No.5, 2014.

[11] Z. Abdullah, A. R. Hamdan, "*Hierarchical Clustering Algorithms in Data Mining*", World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:9, No:10, 2015.