International Journal of Research in Advent Technology, Vol.3, No.7, July 2015
E-ISSN: 2321-9637

Web Document Clustering using Hybrid Approach in Data Mining

Pralhad S. Gamare[1], G. A. Patil[2]
Computer Science & Technology[1,] Computer Science and Engineering[2]
Department of Technology, Kolhapur[1], D. Y. Patil College of Engineering & Technology, Kolhapur[2]
Email: pralhad.gamare@rediffmail.com[1], gasunikita@yahoo.com[2]

Abstract-There is a huge amount of data present on internet. Everyday millions of web pages are added over the web. Hence the procedure to find intended information on the web is becoming very hectic. Use of search engines returns long list of urls and documents. Visiting these urls to extract the required data is still a overwhelming process and it takes lot of time. Therefore there is a need to properly organize documents into groups using clustering. Categorizing similar documents together into clusters will help the users to find useful information quicker, and will allow them to direct their search in the proper direction. Each cluster contains documents that are very similar to each other and very dissimilar to the documents in other clusters.

This paper focuses on Web Document Clustering which uses concept based analysis algorithm along with AHC algorithm to organize the web documents and urls into clusters by considering both the contents of web document and hyperlinks which acts as citation between different web pages.

Index Terms- clustering, web, urls, concept.

1. INTRODUCTION

With the continuous increase in the amount of information loaded onto the internet, information retrieval and organization has become a very crucial problem for the average user. Even with the presence of today's search engines that index the web it is hard to go through the large number of returned documents in a response to a user query. This fact has lead to the need to organize a large set of documents into categories through clustering. It is believed that 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. Clustering is a form of unsupervised classification, which means that the categories into which the collection must be partitioned are not known, and so the clustering process involves the discovering of these categories. Cluster analysis, which deals with the organization of a collection of objects into cohesive groups, can play a very important role towards the achievement of this objective. Clustering is an automatic learning technique aimed at grouping a set of objects into subsets or clusters. The goal is to create clusters that are coherent internally, but substantially different from each other. In other words, objects in the same cluster should be as similar as possible, whereas objects in one cluster should be as dissimilar as possible from objects in the other clusters. Hybrid Approach uses concept-based mining model that analyzes terms on the sentence, document, corpus level and Hierarchical Agglomerative Clustering (HAC) to group the similar documents in clusters and the documents are arranged in hierarchical structure to make easy access of web documents. This approach gives better clustering quality as compared to other document clustering approaches.

1.1. Requirements for document clustering methods

Requirements for document clustering algorithms enable us to design more efficient and robust document clustering solutions [3]. These requirements are described as follows:

1.1.1. *Extraction of Informative Features:* The root of any clustering problem lies in the choice of the most representative set of features describing the underlying data model. The set of extracted features has to be informative enough so that it represents the actual data being analyzed. Also it is important to reduce the number of features because high dimensional feature space always has severe impact on the algorithm scalability.

1.1.2. *Overlapping Cluster Model:* When clustering documents, it is necessary to put those documents in their relevant clusters, which means some documents might belong to more than one cluster. An overlapping cluster model allows this kind of multi-topic document clustering. In some cases it will be

desirable to have disjointed clusters when each document must belong to only one cluster; in these cases one of the non-overlapping clustering algorithms can be used, or a set of disjoint clusters could be generated from fuzzy clustering after de fuzzifying cluster memberships.

1.1.3. *Scalability:* In the web domain, a simple search query might return hundreds, and sometimes thousands, of pages. It is necessary to be able to cluster those results in a reasonable time. An online clustering algorithm should be able to perform the clustering in linear time if possible. An offline clustering algorithm can exceed that limit, but with the merit of being able to produce higher quality clusters.

1.1.4. *Noise Tolerance:* A potential problem faced by many clustering algorithms is the presence of noise and outliers in the data. A good clustering algorithm should be robust enough to handle these types of noise, and produce high quality clusters that are not affected by noise.

1.1.5. *Incremental: T*he web is to be able to update the clusters incrementally. New documents should be added to their respective clusters as they arrive without the need to re-cluster the whole document set. Modified documents should be re-processed and moved to their respective clusters, if applicable.

1.1.6. *Result Presentation:* A clustering algorithm is as good as its ability to present a concise and accurate description of the clusters it produces to the user. The cluster summaries should be representative enough of their respective content, so that the users can determine at a glance which cluster they are interested in.

1.2. Properties of Clustering Algorithms

Following are few properties of Clustering Algorithms [4].

1.2.1. *Data Model:*

1.2.2. *Similarity Measure:*

1.2.3. *Cluster Model:*

2. RELATED WORK

There are many document clustering approaches proposed in the literature. They differ in many parts, such as the types of attributes they use to characterize the documents, the similarity measure used, the representation of the clusters etc. Based on the characteristics or attributes of the documents that are used by the clustering algorithm, the different approaches can be used as follows.

Text based, in which the clustering is based on the content of the document, and link based, based on the link structure of the pages in the collection.

Shady Snehata et al. proposed an efficient concept based mining model for enhancing Text Clustering in paper [1]. This model can efficiently find significant matching concepts between documents, according to the semantics of their sentences. The similarity between documents is calculated based on a new concept-based similarity measure. The proposed similarity measure takes full advantage of using the concept analysis measures on the sentence, document, and corpus levels in calculating the similarity between documents. This approach works well for static collection of documents only and fails for dynamic collection of web pages.

Michael Steinbach et al. Uses K- means algorithm for clustering in paper [11]. Linear time clustering algorithms are the best candidates to comply with the speed requirement of online clustering. These include the K-Means algorithm and the Single-Pass method. One advantage of the K-Means algorithm is that, it can produce overlapping clusters. Its chief disadvantage is that it is known to be most effective when the desired clusters are approximately spherical with respect to the similarity measure used. There is no reason to believe that documents should fall into approximately spherical clusters. The Single-Pass method also suffers from this disadvantage, as well as from being order dependant and from having a tendency to produce large clusters. It is sensitive to input parameters.

Ricardo Campos et al. Proposed automatic hierarchical clustering of web pages in paper [12]. They use architecture called WISE which was composed of four main parts. The selection of relevant pages from the set of all retrieved documents by the search engine, the integration of the SENTA software that extracts phrases from raw texts, the detection of relevant terms that characterize the document using the WEBSPY software that implements the web content mining techniques and the presentation of the documents into a hierarchical structure. Their main contribution was to the field of web content mining techniques applied to the overall information within texts which allows deep semantic analysis of web documents but improvements are necessary especially in terms of merging clusters as data sparseness usually produce too many clusters.

3. PROPOSED ARCHITECTURE

Existing system gives us the several documents that are related to our query. To overcome drawback presents in previous papers, this project proposed a new scheme for web Document Clustering using Hybrid Approach in Data Mining. Our proposed system will provide the related and most relevant documents that user wants or which gives the appropriate documents as a result. The scope of the project is limited to the use of clustering of the web documents using Hybrid Approach such as content as well as hyperlinks using hierarchical agglomerative algorithm and Link based algorithms. The proposed Hybrid Approach uses Concept-Based Mining Model and Hierarchical Agglomerative Clustering(HAC) as a document clustering algorithm along with link based algorithm to cluster the web documents considering both the content of web page as well as and the links of a web page in order to use as much information as possible for the clustering.

Fig. 1 shows our proposed system Architecture that uses the Hybrid Approach (HAC algorithm and Link based algorithm) in order to cluster the documents focusing on both the contents of the web page as well as hyperlinks in the pages.

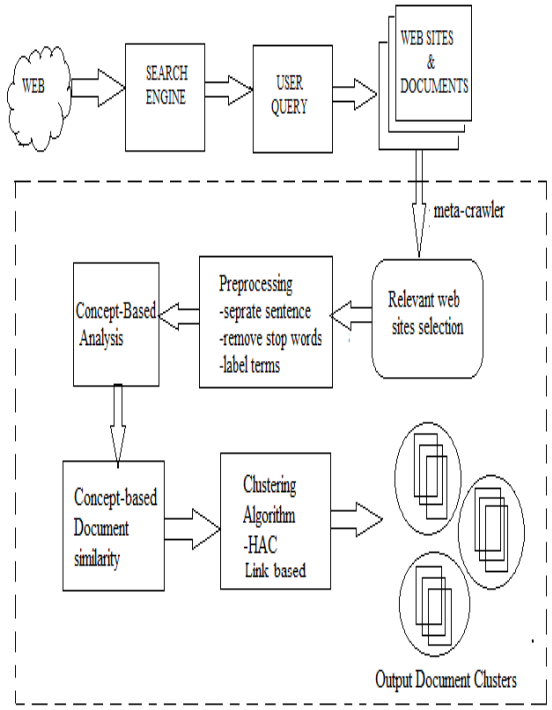

Figure 1. Proposed System

Our proposed work consists of following steps.

1. Retrieve the list of results of the search engine for a given query (meta-crawler).
2. Select the most important results from all the retrieved URLs. For that purpose, we applied a function that chooses from the returned documents, the best ones using following function,
average _ relevance = # returned URLs/# different absolute URLs
3. Now use concept based model to preprocess the documents.
4 Apply concept-based mining model which consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure to achieve an accurate analysis of concepts.
5. Then apply the Concept-Based Document Similarity which allows measuring the importance of each concept with respect to the semantics of the sentence, the topic of the document, and the discrimination among documents in a corpus. Quality of clustering is evaluated using two quality measures, F-measure and Entropy.
6. Use Hierarchical Agglomerative Clustering (HAC) to group the similar documents in clusters and the documents are arranged in hierarchical structure to make easy access of web documents.

3.1. Concept based mining model

A raw text document is the input to the proposed model [1]. Each document has well-defined sentence boundaries. Each sentence in the document is labeled automatically based. After running the semantic role labeler, each sentence in the document might have one or more labeled verb argument structures. The number of generated labeled verb argument structures is entirely dependent on the amount of information in the sentence. The sentence that has many labeled verb argument structures includes many verbs associated with their arguments. The labeled verb argument structures, the output of the role labeling task, are captured and analyzed by the concept-based mining model on sentence, document, and corpus levels. In this model, both the verb and the argument are considered as terms. One term can be an argument to more than one verb in the same sentence. This means that this term can have more than one semantic role in the same sentence. In such cases, this term plays important semantic roles that contribute to the meaning of the sentence. In the concept-based mining model, a labeled terms either word or phrase is considered as concept.

The objective behind the concept-based analysis task is to achieve an accurate analysis of concepts on the sentence, document, and corpus levels rather than a single-term analysis on the document only.

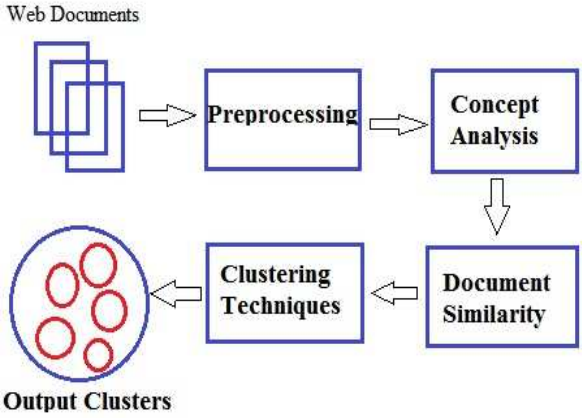

Fig.2 Concept-based mining model system.

Concept-based mining model Algorithm:

1. d_{doci} is a new Document
2. L is an empty List (L is a matched concept list)
3. s_{doci} is a new sentence in d_{doci}
4. Build concepts list C_{doci} from s_{doci}
5. for each concept $c_i \epsilon C_i$ do
6. compute ctf_i of c_i in d_{doci}
7. compute tf_i of c_i in d_{doci}
8. compute df_i of c_i in d_{doci}
9. d_k is seen document, where k –{0; 1; . . . ; doci _ 1}
10. s_k is a sentence in d_k
11. Build concepts list C_k from s_k
12. for each concept $c_j \epsilon C_k$ do
13. if (c_i == c_j) then
14. update df_i of c_i
15. compute ctfweight – avg (ctf_i ; ctf_j)
16. add new concept matches to L
17. end if
18. end for
19. end for
20. output the matched concepts list L

3.2 Concept-Based Document Similarity:

The concept-based similarity measure relies on three critical aspects. First, the analyzed labelled terms are the concepts that capture the semantic structure of each sentence. Second, the frequency of a concept is used to measure the contribution of the concept to the meaning of the sentence, as well as to the main topics of the document. Last, the number of documents that contains the analyzed concepts is used to discriminate among documents in calculating the similarity. These aspects are measured by the proposed concept-based similarity measure which measures the importance of each concept at the sentence level by the ctf measure, document level by the tf measure, and corpus level by the df measure. This similarity measure is a function of the following factors.

Algorithm:

1. The number of matching concepts, m, in the verb argument structures in each document d,
2. The total number of sentences, sn, that contain matching concept c_i in each document d,
3. The total number of the labeled verb argument structures, v, in each sentence s,
4. The ctf_i of each concept c_i in s for each document d, where i - 1; 2; . . .; m, as mentioned in Sections 4.1.1.1 and 4.1.1.2,
5. The tf_i of each concept c_i in each document d, where i - 1; 2; . . .;m.
6. The df_i of each concept ci, where i - 1; 2; . . .;m,
7. The length, l, of each concept in the verb argument structure in each document d,
8. The length, Lv, of each verb argument structure which contains a matched concept, and
9. The total number of documents, N, in the corpus.

3.3 AHC Algorithm

AHC algorithm is used to cluster the web documents in hierarchical order. The quality of output cluster is evaluated by using F measure and Entropy.
The hierarchical agglomerative clustering methods:
The hierarchical agglomerative clustering methods differ in the way they calculate the similarity between two clusters. The existing methods are the following:

3.3.1 Single link:

The similarity between a pair of clusters is calculated as the similarity between the two most similar documents, one of which is in each cluster. This method tends to produce long, loosely bound clusters with little internal cohesion (chaining effect). The single link method incorporates useful mathematical properties and can have small computational complexity

3.3.2 Complete link:

The similarity between a pair of clusters is taken to be the similarity between the least similar documents, one of which is in each cluster. This definition is much stricter than that of the single link method and, thus, the clusters are small and tightly bound.

Link-based clustering:

Text-based clustering approaches were developed for use in small, static and homogeneous collections of documents. But www is a huge collection of heterogeneous and interconnected web pages. The link-based document clustering approaches take into account information extracted by the link structure of the collection. The underlying idea is that when two documents are connected via a link there exists a semantic relationship between them, which can be the

basis for the partitioning of the collection into clusters.

Hybrid Algorithm

The steps of the typical AHC algorithm are the following:

1. Assign each document to a single cluster.
2. Compute the similarity between all pairs of clusters and store the result in a similarity matrix, in which the ijth entry stores the similarity between the ith and jth cluster.
3. Merge the two most similar (closest) clusters.

4. RESULT ANALYSIS

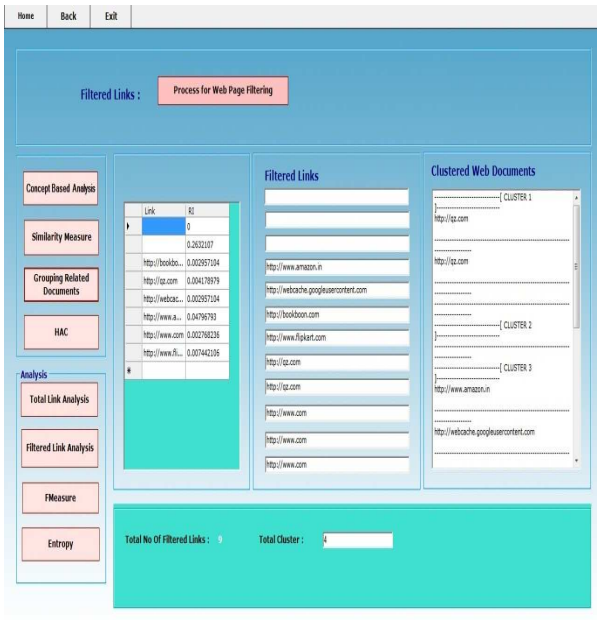

Fig.3 Process for displaying clustered web documents

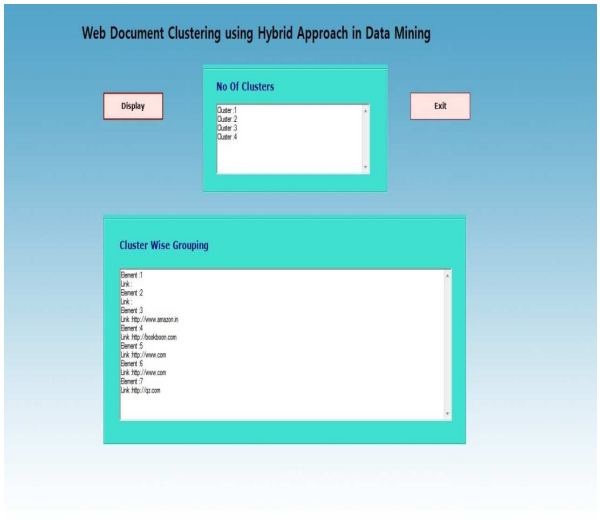

Fig. 4 Output of HAC algorithm

Fig. 5 Total link analysis of query term

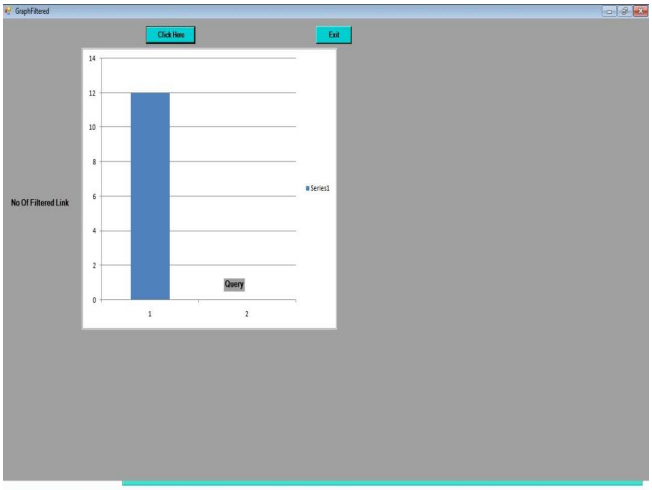

Fig. 6 Filtered link analysis of query term

5. CONCLUSION AND FUTURE WORK

Clustering is a very complex procedure which depends on the data on which it is performed and selection of various parameter values. Hence, a careful selection of these is very crucial. Clustering is a very useful technique which helps to organize and retrieve useful data or information across internet. Web document clustering using hybrid approach gives clusters with semantically identical objects. It makes use of conceptual term frequency, term frequency and document frequency to effectively correlate web documents in to clusters. Clustering can increase the efficiency and the effectiveness of information retrieval. Thus document clustering is very useful to retrieve information application in order to reduce the consuming time and get high precision and recall.

Clustering is also useful in extracting salient features of related Web documents to automatically formulate queries and search for other similar documents on the Web. Use of link-based clustering approaches has proved to be very useful source of information for the clustering process. This work can be extended further for text clustering and excel documents clustering. There are still some challenges for further research. These include the achievement of better quality-complexity tradeoffs, as well as effort to deal with each method's disadvantages. In addition, another very important issue is Incrementality, because the web pages change very frequently and because new pages are always added to the web. Also, the fact that very often a web page relates to more than one subject should also be considered and lead to algorithms that allow for overlapping clusters.

REFERENCES

[1] Shady Shehata, Member, IEEE, Fakhri Karray, Senior Member, IEEE and Mohamed S. Kamel, Fellow, IEEE, *"An Efficient Concept-Based Mining Model for Enhancing Text Clustering"* IEEE transactions on Knowledge and Data Engineering, Vol.22, No.10, October 2010.

[2] O. Zamir and O. Etzioni," *Web Document Clustering: A Feasibility Demonstration*," Proc. of the 21st ACM SIGIR Conference, pp 46-54, University of Washington.

[3] N. Oikonomakou, M. Vazirgiannis," *A Review of Web Document Clustering Approaches*" Athens University of Economics & Business Greece.

[4] K. Cios, W. Pedrycs, R. Swiniarski, "*Data Mining – Methods for Knowledge Discovery"*, Kluwer Academic Publishers, 1998.

[5] R. Krishnapuram, A. Joshi, L. Yi, *"A Fuzzy Relative of the k-Medoids Algorithm with Application to Web Document and Snippet Clustering"*, Proc. *IEEE Intl. Conf. Fuzzy Systems*, Korea, August 1999.

[6] Khaled M. Hammouda, "Web Mining: Clustering Web Documents A Preliminary Review", Department of Systems Design Engineering, University of Waterloo, Feb. 2001

[7] Z. Jiang, A. Joshi, R. Krishnapuram, L. Yi, *Retriever:" Improving Web Search Engine Results Using Clustering*," Technical Report, CSEE Department, UMBC, 2000.

[8] A. Strehl, J. Ghosh, and R. Mooney." Impact of Similarity Measures onWeb-Page Clustering." In *Workshop for Artificial Intelligence for Web Search*, July 2000.

[9] D. Crabtree, X. Gao, and P. Andreae." Improving web clustering by cluster selection". In The 2005 IEEE/WIC/ACM International Conference on Web Intelligence, pages 172–178, September 2005.

[10] Kate A. Smith and Alan Ng. "Web page clustering using a self-organizing map of user navigation patterns". *Decision Support Systems*, 35(2):245–256, 2003.

[11] M. Steinbach, G. Karypis, V. Kumar, "*A Comparison of Document Clustering Techniques*" TextMining Workshop, KDD, 2000.

[12] Ricardo Campos, "*Automatic Hierarchical Clustering of Web Pages*", Centre of Human Language Technology and Bioinformatics, University of Beira Interior.

[13] Y. W. Wong and A. Fu, "*Incremental Document Clustering for Web Page Classification"*, Int. Conf. on Info. Society in the 21st century: emerging technologies and new challenges (IS2000), Nov 5-8, 2000, Japan.

[14] G. Karypis, E. Han, V. Kumar, *CHAMELEON:" A Hierarchical Clustering AlgorithmUsing Dynamic Modeling"*, IEEE Computer 32, pp. 68-75, August 1999.

[15] P. Mitra, C. Murthy, and S.K. Pal, "Unsupervised Feature Selection Using Feature Similarity," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 301-312, Mar. 2002.

[16] M. Junker, M. Sintek, and M. Rinck, "Learning for Text Categorization and Information Extraction with ILP," Proc. First Workshop Learning Language in Logic, 1999.

[17] Roger Pressman, Software Engineering: A Practitioners Approach, (6th Edition), McGraw Hill, 2010

[18] Y. Yang and J. Pedersen, "*A Comparative Study on Feature Selection in Text Categorization"*, In Proc. of the 14th International Conference on Machine Learning, pages 412-420, Nashville, TN, 1997.

[19] Hinrich Schutze and Craig Silverstein," *Projections for Efficient Document Clustering"*, SIGIR '97, Philadelphia, PA, 1997.

[20] Oren Zamir, Oren Etzioni, Omid Madani, Richard M. Karp, "*Fast and Intuitive Clustering of Web Documents"*, KDD '97, Pages 287-290, 1997.

[21] R. Krishnapuram, A. Joshi, L. Yi," *A Fuzzy Relative of the k-Medoids Algorithm with Application to Web Document and Snippet Clustering"*, Proc. *IEEE Intl. Conf. Fuzzy Systems*, Korea, August 1999.