International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 05 | August-2015 www.irjet.net p-ISSN: 2395-0072

Efficient Clustering of Web Documents Using Hybrid Approach in Data Mining

Pralhad Sudam Gamare[1], Ganpati A. Patil[2]

[1] P.G. Student, Computer Science and Technology, Department of Technology-Shivaji University-Kolhapur, Maharashtra, India
[2] Associate Professor, Computer Science and Engineering, D. Y. Patil College of Engineering and Technology, Kolhapur, Maharashtra, India
[1] pralhad.gamare@rediffmail.com, [2] gasunikita@yahoo.com

Abstract - ***There is a huge amount of data present on internet. The procedure to find important information on the web can be very hectic. Using today's search engines it is hard to go through the big number of returned urls and documents. Hence there is a need to arrange documents into groups using clustering. Categorizing related 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. Cluster analysis helps to organize set of objects into cohesive groups and can leads to the achievement of this objective. Clustering is an automatic learning technique aimed at grouping a set of documents into groups called clusters. The intention is to form clusters that are coherently similar, but substantially different from one another. Documents belonging to one cluster are similar to each other and are very dissimilar to the documents in other cluster. This paper focuses on web page clustering which uses concept analysis and HAC algorithm to produce high quality clusters by considering contents and hyperlinks of the web page.***

Key Words: *Clustering, Concept, HAC algorithms, urls*

1. INTRODUCTION

Nowadays, the internet has become the largest data repository, facing the problem of information overload. Evidently there is a tremendous proliferation in the amount of information found today on the largest shared information source, the World Wide Web (or simply the Web).The existence of an abundance of information, in combination with the dynamic and heterogeneous nature of the web, makes information retrieval a tedious process for the average user. Even with the presence of today's search engines that index the web it is hard to wade 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 (due to a user query or simply a collection 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. Usually, a user searching for information submits a query composed by a few keywords to a search engine (such as Google (http://www.google.com)). The search engine performs exact matching between the query terms and the keywords that characterize each web page and presents the results to the user. These results are long lists of URLs, which are very hard to search. Furthermore, users without domain expertise are not familiar with the appropriate terminology thus not submitting the right (in terms of relevance or specialization) query terms, leading to the retrieval of more irrelevant pages. This has led to the need for the development of new techniques to assist users effectively navigate, trace and organize the available web documents, with the ultimate goal of finding those best matching their needs. One of the techniques that can play an important role towards the achievement of this objective is document clustering. 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. Clustering can be used as a very powerful mechanism for browsing a collection of documents or for presenting the results of the retrieval. A typical retrieval on the Internet will return a long list of web pages. The organization and presentation of the pages in small and meaningful groups (usually followed by short descriptions or summaries of the contents of each group) gives the user the possibility to focus exactly on the subject of his interest and find the desired documents more quickly. Furthermore, the presentation of the search results in clusters can provide an overview of the major subject areas related to the user's topic of interest. Document clustering has been investigated for use in a number of different areas of text mining and information retrieval. Initially, document clustering was investigated for improving the precision or recall in information retrieval systems and as an efficient way of finding the nearest neighbors of a document. More recently, clustering

has been proposed for use in browsing a collection of documents or in organizing the results returned by a search engine in response to a user's query. Document clustering has also been used to automatically generate hierarchical clusters of documents.

2. LITERATURE SURVEY

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.

a) Text based, in which the clustering is based on the content of the document,
b) Link based, based on the link structure of the pages in the collection.

Most algorithms in the first category were developed for use in static collections of documents that were stored and could be retrieved from a database and not for collections of web pages. Various text based algorithms are partitional, hierarchical, graph based, neural network-based and probabilistic each having their own advantages and disadvantages. The most widely used document clustering algorithms falls in two categories: partitional and hierarchical. K-means is also used but it is very sensitive to input parameters.

But World Wide Web is a *directed graph*. This means that apart from its content, a web page contains other characteristics that can be very useful to clustering. The most important among these are the hyperlinks that play the role of citations between the web pages. The basic idea is that when two documents are cited together by many other documents (i.e. have many common incoming links) or cite the same documents (i.e. have many common outgoing links) there exists a semantic relationship between them. In the Web Information Retrieval literature there are many applications based on the use of hyperlinks in the clustering process and the calculation of the similarity based on the link structure of the documents has proven to produce high quality clusters.

Here Suffix Tree Clustering algorithm is very usually used but instead of whole web document it only works on Snippets. Snippets may not be a good description of a web page and Snippets usually introduce noise. Hence this method is also having drawbacks.

3. RELATED WORK

3.1 Partitional Algorithms:

Partitional Clustering Algorithms creates the clusters in one step as opposed to several steps. Only one step of clusters is created, although several different sets of clusters may be created internally within the various algorithms. Since only one set of clusters is output, the user must input the desired number, k, of clusters. in addition, some metric or criterion function is used to determine the goodness of any proposed solution.

a) Minimum Spanning Tree

This is a very simplistic approach, but it illustrates how partitional algorithm works. Since the clustering problem is to define a mapping, the output of this algorithm shows the clusters as a set of ordered pairs (t_i, j) where $f(t_i)=K_j$.

b) Squared Error Clustering Algorithm

Squared Error Clustering Algorithm minimizes the squared error. The squared error for a cluster is the sum of the squared Euclidean distances between each element in the cluster and the cluster centroid, C_k . given a cluster K_i, let the set of items mapped to that cluster be$\{t_{i1},t_{i2},\ldots,t_{im}\}$. The squared error is defined as

$Se_{ki}=\sum_{j=1}^{m}=\| t_{ij}-C_k \|^2$

Given a set of clusters $K=\{K_1,K_2,K_3,\ldots,K_k\}$, the squared error for K is defined as

$Se_k=\sum_{j=1}^{m} Se_{kj}$

c) K-Means Clustering

K-Means is an iterative clustering algorithm in which items are moved among set of clusters until the desired set is reached. A high degree of similarity among elements in clusters is obtained, while a high degree of dissimilarity among elements in different clusters is achieved simultaneously. The cluster mean of $K_i=\{t_{i1},t_{i2},\ldots,t_{im}\}$is defined as, $m_i=1/m(\sum_{j=1}^{m}t_{ij})$ The K-means Algorithm assumes that the desired number of clusters, k, is an input parameter.

d) Nearest Neighbor Algorithm

With Nearest Neighbor Algorithm, items are iteratively merged into the existing clusters that are closest. Here a threshold, t, is used to determine if items will be added to existing clusters or if a new cluster is created. Its complexity depends on the number of items. For each loop, each item must be compared to each item already in a cluster.

3.2 Large Database Clustering Algorithms

When clustering is used with dynamic databases, the above algorithms may not be appropriate. It has been argued that to perform effectively on large databases, a clustering algorithm should require no more than one scan of the database, be suspend able, stoppable and resumable, be able to update the results incrementally as data are added or removed from the database, work with limited main memory, process each tuple only once.

a) BIRCH Algorithm

BIRCH (balanced iterative reducing and clustering using hierarchies) assumes that there may be a limited main memory and achieves a linear I/O time requiring only one database scan. Here a tree is built that captures needed information to perform clustering. Clustering is then performed on the tree itself, where labeling of nodes in the tree contain required data to calculate distance values. It uses clustering feature, which is a triple that contains information about a cluster.

b) DBSCAN Algorithm

DBSCAN (density-based spatial clustering of applications with noise) is used to create clusters with a minimum size and density. Density is a minimum number of points within a certain distance of each other. It ensures that second point is "close enough" to the first point. Also core points must be close enough to each other. These core points form the main portion of a cluster in that they are all close to each other.

c) CURE Algorithm

CURE(Clustering Using REpresentatives) algorithm is used to handle outliers. First a constant number of points, c, are chosen from each cluster. These well scattered points are then shrunk towards the clusters centroid by applying a shrinkage factor, α. When α is 1, all points are shrunk to just one-the centroid. At each step in the agglomerative algorithm, clusters with the closest pair of representative points are chosen to be merged.

4. PROPOSED WORK

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.

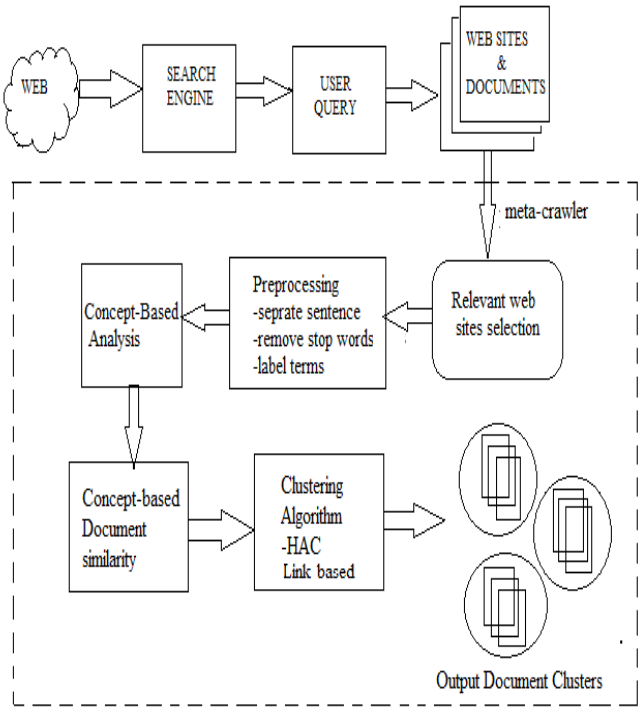

Fig -1: Proposed System Architecture

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. This approach will give better clustering quality as compared to other document clustering algorithms.

Concept Based Mining Model:
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 [1].

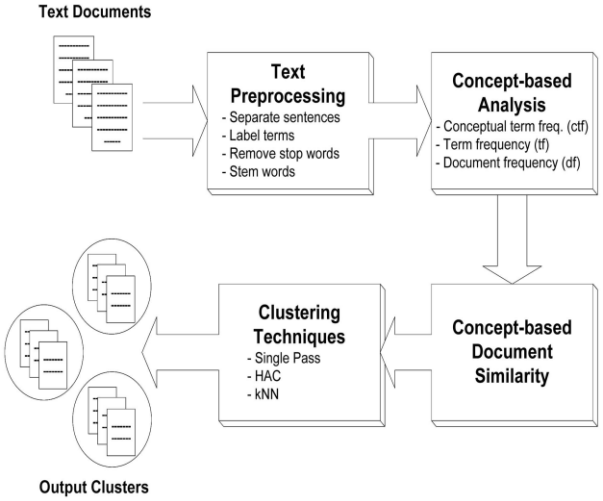

Fig – 2: Concept Based Mining Model

Sentence-Based Concept Analysis-
To analyze each concept at the sentence level, a new concept-based frequency measure, called the conceptual term frequency (ctf) is proposed. The ctf calculations of concept c in sentence s and document d are as follows:

1. Calculating ctf of Concept c in Sentence s
The ctf is the number of occurrences of concept c in verb argument structures of sentences. The concept c, which frequently appears in different verb argument structures of the same sentence s, has the principal role of contributing to the meaning of s. In this case, the ctf is a local measure on the sentence level.

2. Calculating ctf of Concept c in Document d
A concept c can have many ctf values in different sentences in the same document. Taking the average of the ctf values of concept c in its sentences of document d measures the overall importance of concept c to the meaning of its sentences in document d. A concept, which has ctf values in most of the sentences in a document, has a major contribution to the meaning of its sentences that leads to discover the topic of the document. Thus, calculating the average of the ctf values measures the overall importance of each concept to the semantics of a document through the sentences.

3. Document-Based Concept Analysis
To analyze each concept at the document level, the concept based term frequency tf, the number of occurrences of a concept (word or phrase) c in the original document, is calculated. The tf is a local measure on the document level.

4. Corpus-Based Concept Analysis
To extract concepts that can discriminate between documents, the concept-based document frequency df, the number of documents containing concept c, is calculated. The df is a global measure on the corpus level. This measure is used to reward the concepts that only appear in a small number of documents as these concepts can discriminate their documents among others. The process of calculating ctf, tf, and df measures in a corpus is attained by the proposed algorithm which is called Concept-based Analysis Algorithm.

5. CONCLUSION

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. This work can be extended further for text clustering and excel documents clustering. 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. Use of link-based clustering approaches has proved to be very useful source of information for the clustering process. 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. Finally, more attention should also be given to the description of the clusters' contents to the users, the labeling issue.

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.

[3] A. Strehl, J. Ghosh, and R. Mooney." Impact of Similarity Measures on Web-Page Clustering." In Workshop for Artificial Intelligence for Web Search, July 2000.
[4] 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.
[5] 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.
[6] Hinrich Schutze and Craig Silverstein," Projections for Efficient Document Clustering", SIGIR '97, Philadelphia, PA, 1997.
[7] Oren Zamir, Oren Etzioni, Omid Madani, Richard M. Karp, "Fast and Intuitive Clustering of Web Documents", KDD '97, Pages 287-290, 1997.
[8] 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.
[9] A. K. Jain and R. C. Dubes," Algorithms for Clustering Data", John Wiley & Sons, 1988.
[10]Z .Jiang, A. Joshi, R. Krishnapuram, L. Yi, Retriever:" Improving Web Search Engine Results Using Clustering," Technical Report, CSEE Department, UMBC, 2000.

BIOGRAPHIES

Mr. Pralhad Sudam Gamare is working as Assistant Professor in Department of Computer Engineering at RMCET Ambav, Mumbai University since July 2006. He has completed his B.E. (CE) from Mumbai University and currently persuing his MTECH from Shivaji University, Kolhapur, Maharashtra, India.

Prof. G. A. Patil is working as Associate Professor and HOD at D.Y. Patil College of Engineering and Technology, Kolhapur. He is having teaching experience of 24 years and currently persuing his PhD in Computer Science & Engineering.