Design and Comparison of Agglomerative Hierarchical Clustering

Sarika
Department of Computer Engineering
MIET Institute of Engineering Meerut, UP

Mukesh Rawat, PhD
Department of Computer Engineering
MIET Institute of engineering Meerut, UP

ABSTRACT
As more and more documents are available in the form of hypertext in world wide web, proper clustering of documents are required for generating the similar results fetched by the search engine specific to a user entered search query as the documents within the cluster are similar to each other. In Agglomerative approach of clustering the clusters of documents are merged into cluster unless until all the clusters belong to a root cluster .In this paper design and comparison of two agglomerative hierarchical clustering of documents is done on parameters such as cluster generation, relevance of result get against a query and purity of clusters.

Keywords
Agglomerative, relevance, purity, cluster, search engine.

1. INTRODUCTION
Clustering calculations amass an arrangement of records into subsets or groups. The calculations objective is to make clusters that are cognizant inside, however obviously unique in relation to each other. At the end of the day, reports inside a cluster ought to be as comparable as could be expected under the circumstances; and archives in one cluster ought to be as unique as could be expected under the circumstances from records in different clusters. Clustering can be viewed as the most vital unsupervised learning issue; along these lines, as each other issue of this kind, it manages finding a structure in an accumulation of unlabeled information. Clustering is unsupervised learning since it doesn't utilize predefined classification names connected with information things .Clustering calculations is utilized as a part of separating valuable data in vast database. Clustering calculations are designed to discover structure in the present information, not to classes future information. The objective of clustering is to arrange information by discovering some "sensible" gath. Clustering is the most widely recognized type of unsupervised learning. No super-learning vision implies that there is no human master who has doled out archive to classes. In grouping, it is the circulation and cosmetics of the information that will decide group participation. Clustering can likewise accelerate look. Grouping is a numerical instrument that endeavors to find structures or certain examples in a dataset, where the items inside every group demonstrate a specific level of closeness. Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc. The classic example of this is species taxonomy. Gene expression data might also exhibit this hierarchical quality (e.g. neurotransmitter gene families). Agglomerative hierarchical clustering starts with every single object (gene or sample) in a single cluster. Then, in each successive iteration, it agglomerates (merges) the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster. In Agglomerative approach of clustering the clusters of documents are merged into cluster unless until all the clusters belong to a root cluster .In this paper design and comparison of two agglomerative hierarchical clustering of documents is done on parameters such as cluster generation, relevance of result get against a query and purity of clusters.

2. DESIGN OF AGGLOMERATIVE HIERARCHICAL CLUSTERING MODEL

2.1 Preprocessing
Most of the documents are available in the form of hyper text and the textual information is available between the tags, the problem is to filter out the text from tags.

***Solution**: The solution is to use html to text parser which removes tags and filter out the text. Next is to remove unnecessary words from the text which contribute no meaning to the document.*

First of all we are having collection of 100 documents. All these documents are in HTML form .We will change these HTML files into Text files ,for changing in text files we will use HTML2 TEXT PARSER. By using this html2 text parser in preprocessing we will change these html files into text files. Now we are having the Text Repository. Now on this text repository we will apply the Agglomerative Hierarchical Clustering and Jaccard coefficient and will cluster the documents.

2.2 Hierarchical Automatic Cluster Generation
First of all we are having collection of text documents. All these documents are in HTML form. We need to convert all these documents in text form.So we have to perform preprocessing for converting all these documents in text form. We need to use html 2 text parser. Now we are having the repository of text documents. Now we will proceed by taking every term one by one. First of all we will take terms of first document. Terms with highest similarity will form new cluster. We will do this for each and every term. We will match each and every term of incoming document with the terms of generated cluster documents. If the document terms matches with the documents of generated cluster then assign the documents to the cluster. If the document terms do not match with the documents of generated cluster then generate new cluster. This task is repeated again and again until the matching of all the documents has been done properly .By performing all this we are having clusters with high intra cluster similarity and with low inter similarity. This is beneficial because it is completely scalable. We can add documents to clusters afterwards .We need not to define the number of clusters at prior level. Clustering takes a special place since it is reliable and easy to configure the clusters. This also reduces the searching time because similar documents are there wihin a single cluster. This method also

provides efficient and effective processing of documents. Efficient processing implies minimizing the amount of time and space required to access information, whereas effective processing means identifying accurately which information is relevant to the user. Traditionally, efficiency and effectiveness are at opposite ends and hence the challenge of to find the ways to create a balance between efficient and effective data processing..

Hierarchical algorithms are more versatile than partitional algorithms. For example, the single-link clustering algorithm works well on data sets containing non-isotropic clusters including well-separated, chain-like, and concentric clusters, whereas a typical partitional algorithm such as the *k*-means algorithm works well only on data sets having isotropic clusters .On the other hand, the time and space complexities of the partitional algorithms are typically lower than those of the hierarchical algorithms. Proper clustering of web documents is required for efficient searching of documents according to the terms .Single level cluster generation of web documents generates the large set of clusters and the searching requires much more time .So, Hierarchical agglomerative clustering of web documents is suggested which merge the documents generated by flat clustering into a one cluster in a hierarchical form from bottom to top. Clustering is the most common form of unsupervised learning. No super-vision means that there is no human expert who has assigned document classes. The main goal of this clustering is to enable users to extract information from textual resources. How the documented can be proper annotated, presented and classified, so the documents categorization consist several challenges, proper annotation to the documents, appropriate document representation, an appropriate classifier function to obtain good generalization. We have applied single linkage and complete linkage hierarchical agglomerative clustering method for getting efficient retrieval model. The objective of clustering is to discover brilliant clusters with the end goal that the between group likeness is low and the intra-group similitude is high.

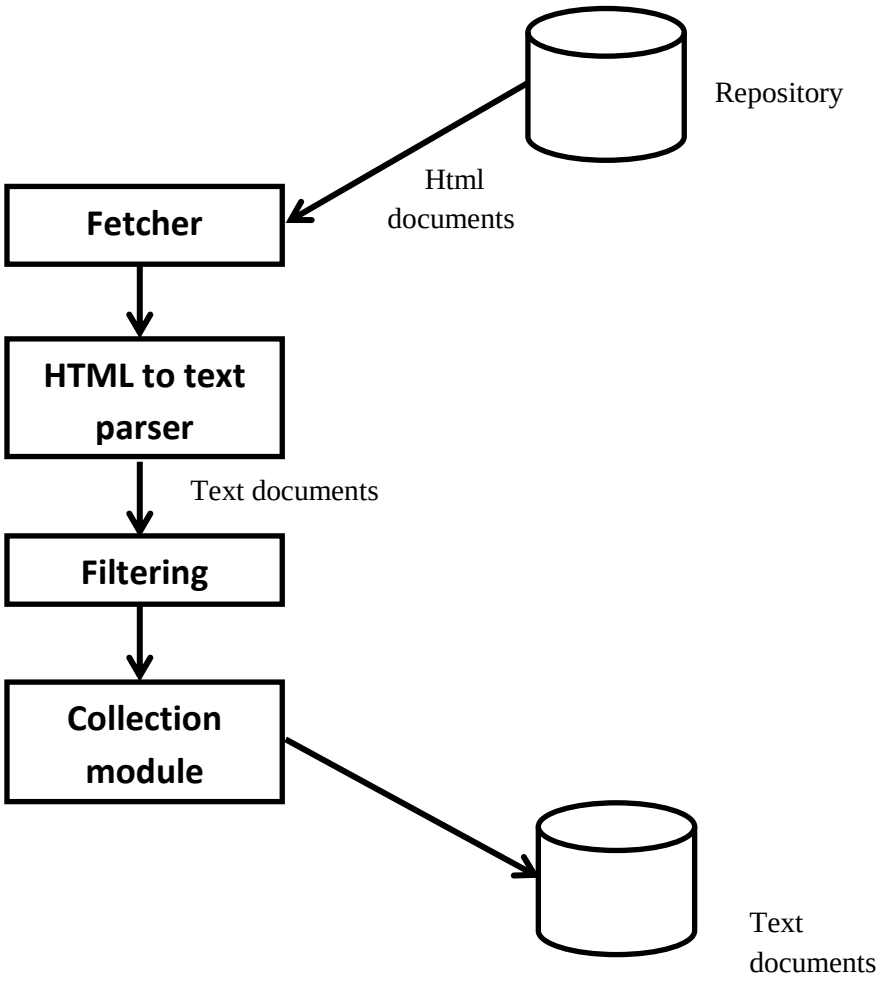

Fig 1: Proposed Hierarchical Clustering model

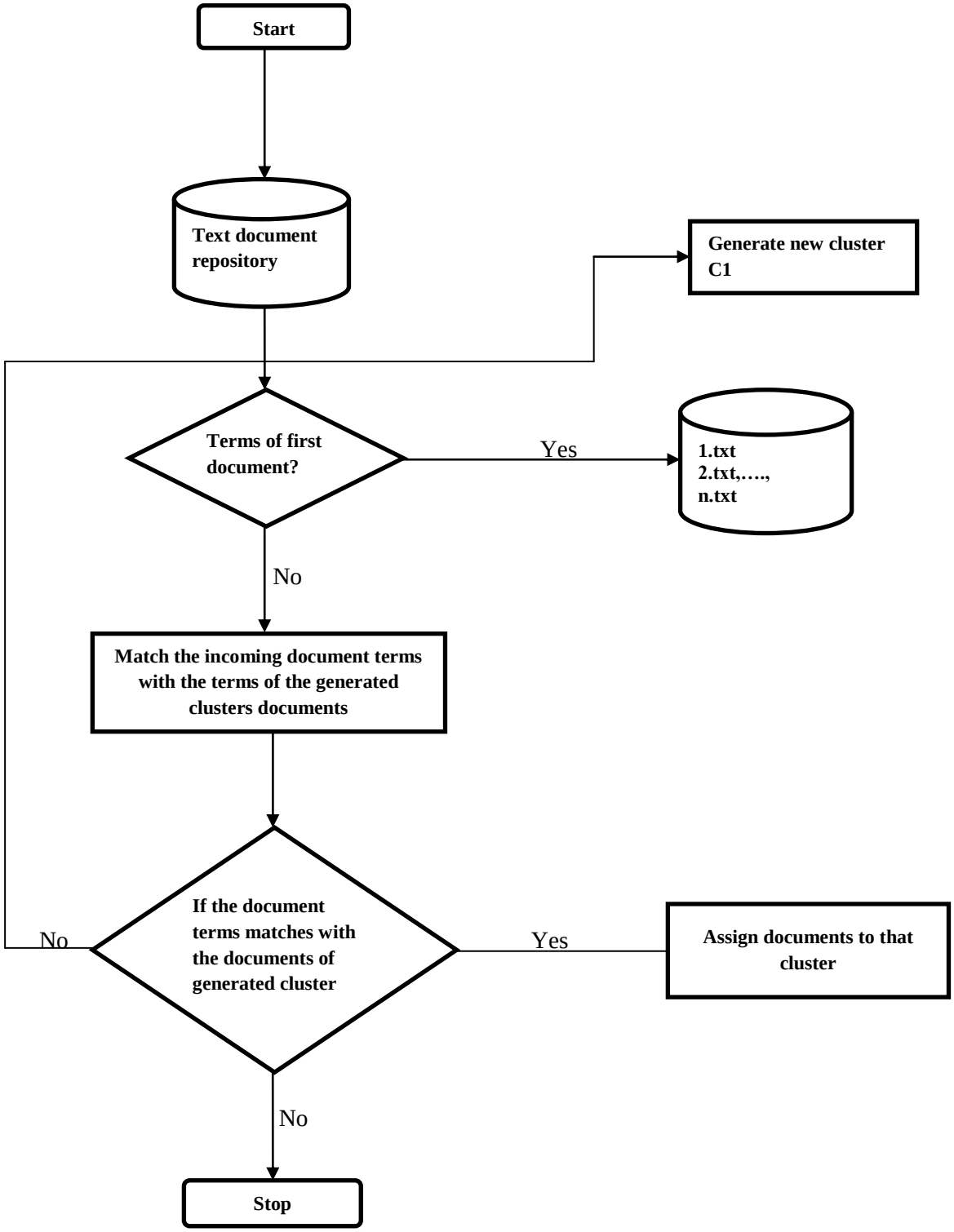

Fig 2: Hierarchical Automatic Cluster Generation

3. CLUSTER GENERATION BY SINGLE LINKAGE AND COMPLETE LINKAGE

Here we are having five clusters. Cluster C1 contains different documents like D1, D2, D3 upto Dk. In the same way cluster C2 contains different documents like P1, P2, P3 up to pk. Cluster C3 contains documents M1, M2, M3 up to mk. Cluster C4 contains documents N1, N2, N3 up to nk and Cluster C5 contains documents I1, I2, I3upto ik. This representation is at the ground level.

3.1 First Level Cluster Generation by Single Linkage

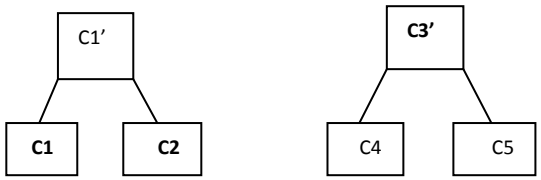

Fig 3: Level 1 by Single Linkage

Here the generation of new cluster is done, the documents of cluster C1 is having similarity with the documents of cluster C2 so these clusters will be merged and will form a new cluster C1' at first level.In the same way the documents of cluster C4 is having similarity with the documents of cluster C5 so they will be merged and will form a new cluster C3'.

3.2 First level cluster generation by Complete Linkage

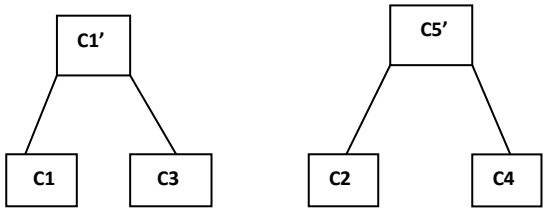

Fig 4: Level 1 by Complete Linkage

3.3 Second level Cluster Generation by Complete linkage

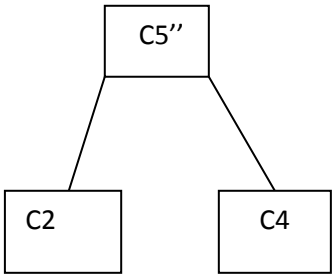

Fig 5 : Level 2 by Complete Linkage

4. RESULT ANALYSIS

The result is measured via three metrics: precision, recall, and F-measure. Precision is the percentage of correctly identified document over all the documents in the repository, while Recall is the percentage of correctly identified document over all the correctly identified document and unidentified document. F-measure incorporates both precision and recall.

Here the generation of new cluster is done, the documents of cluster C1 is having similarity with the documents of cluster C3 according to the criteria of complete linkage so these clusters will be merged and will form a new cluster C1'at first level. In the same way the documents of cluster C2 is having similarity with the documents of cluster C4 so they will be merged and will form a new cluster C5'.

Table 1: Showing P , R , F measure of documents

No. of Documents	Search Term	C	W	M	P	R	F
100	Deadlock	60	6	20	0.909091	0.75	0.81318
150	Scheduling	65	6	23	0.915493	0.738636	0.81761
200	Process termination	70	8	12	0.897436	0.853659	0.875
250	SJF	74	6	11	0.925	0.870588	0.89697
300	Memory Management	100	7	22	0.934579	0.819672	0.89536
350	RR	120	8	12	0.9375	0.909091	0.923077
400	FCFS	189	7	23	0.964286	0.891509	0.926471

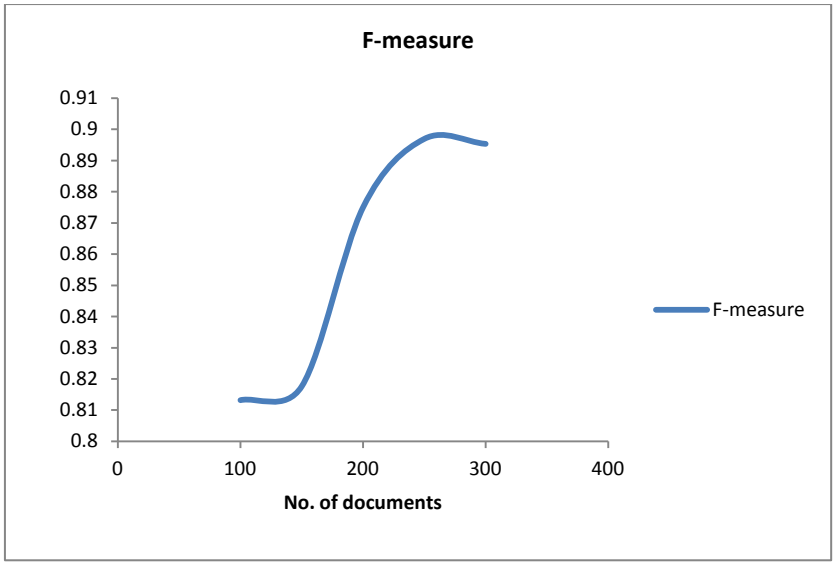

Fig 6: Graphical representation of result analysis

5. CONCLUSION

Compared to traditional clustering, where whole documents are usually clustered on the basis of some physical properties, we have clustered the documents here on the basis of similarity measures and distance between the documents of different clusters. Here we have done the comparison between single linkage and complete linkage clustering methods. The aim of these type of clustering is to minimize the searching time of documents in different clusters. Efficient processing means reducing the amount of time and space required to access information, whereas effective processing means identifying accurately which information is relevant to the user. Traditionally, efficiency and effectiveness are at different ends and hence the challenge to find ways to create a balance between efficient and effective processing.

6. REFERENCES

[1] Nicholas O. Andrews and Edward A. Fox, "Recent Developments in Document Clustering", thesis, October 16, 2007.

[2] Jain and R. Dubes. "Algorithms for Clustering Data." Prentice Hall, 1988.

[3] Chris Staff: Bookmark Category Web Page Classification Using Four Indexing and Clustering Approaches. AH 2008:345-348.

[4] Han J., Kamber M.,"Data Mining: Concepts and Techniques," Morgan Kaufmann (Elsevier),2006.

[5] Seung-sikh, "Keyword based document clustering", report, school of cs, kookim university. Seoul, Korea.

[6] Swatantra Kumar Sahu*, "Classification of Document clustering Approaches", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 2, Issue 5, May 2012.

[7] Charu C. Aggarwal, "A Survey of Text Clustering Algorithms", report, *IBM T. J. Watson Research Center Yorktown Heights, NY.*

[8] Anna Huang, "Similarity Measures for Text Document Clustering", report, Department of Computer Science, The University of Waikato, Hamilton, NewZealand.