ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.6

A Review of Hierarchal Clustering Algorithm and Evaluation

Ekta Chauhan[1], Dr. Amit Asthana[2]
M.Tech Student[1], HOD[2]
Department of CSE
Shubharti University, Meerut, India

Abstract:
Users of Web search engines are often forced to sift through the long ordered list of document "snippets" returned by the engines. The IR community has explored document clustering as an alternative method of organizing retrieval results, but clustering has yet to be deployed on the major search engines. Document clustering is a subset of the larger data clustering, which carries concepts from the fields of information retrieval (IR), natural language processing, and machine learning, among others. Document clustering will hereafter be simply referred to as clustering. Document clustering has been investigated in different areas of text mining and information retrieval. Document clustering has been studied deeply because of its wide application in areas such as Web Mining, Search Engine and Information Retrieval. Document clustering is the automatic organization of documents into clusters or groups, so that, documents within a cluster have high similarity in comparison to one another, but are very dissimilar to documents in other clusters. In other words, the grouping is based on the principle of maximizing intracluster similarity and minimizing inter-cluster similarity. Clustering can also speed up search. Search in the vector space model amounts to finding the nearest neighbors to the query. The inverted index supports fast nearest-neighbor search for the standard IR setting. However, sometimes an inverted index is not used efficiently, e.g., in latent semantic indexing. In such cases, the similarity of the query to every document is calculated, but this is slow. The cluster hypothesis offers an alternative: Find the clusters that are closest to the query and only consider documents from these clusters. Within this much smaller set, we can compute similarities exhaustively and rank documents in the usual way. Since there are many fewer clusters than documents, finding the closest cluster is fast; and since the documents matching a query are all similar to each other, they tend to be in the same clusters. This paper presents various clustering techniques and the cluster purity check mechanism.

Keywords: search engine, Information Retrieval, indexes, Boolean retrieval.

1. INTRODUCTION

Document clustering (or Text clustering) is automatic document organization, topic extraction and fast information retrieval or filtering. A web search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories, as is achieved by Enterprise Search engines. Clustering is a widely adopted technique aimed at dividing a collection of data into disjoint groups of homogenous elements.

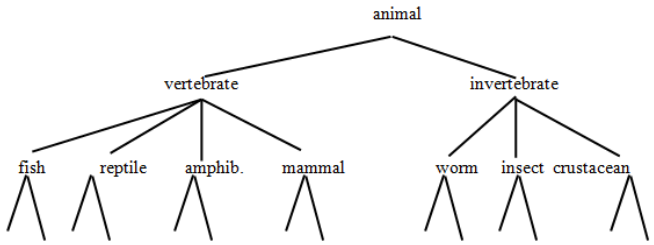
Document clustering [3] has been widely investigated as a technique to improve effectiveness and efficiency in information retrieval. Clustering algorithms attempt to group together the documents based on their similarities. Thus documents relating to a certain topic will hopefully be placed in a single cluster. So if the documents are clustered, comparisons of the documents against the user's query are only needed with certain clusters and not with the whole collection of documents. The fast information retrieval can be further achieved by hierarchical clustering in which the similar clusters are merged together to form higher levels of clustering. Document clustering involves the use of descriptors and descriptor extraction. Document clustering is generally considered to be a centralized process. Clustering is the most common form of unsupervised learning.

Uses of clustering of documents –

- If a collection is well clustered, we can search only the cluster that will contain relevant documents.
- Searching a smaller collection should improve effectiveness and efficiency.

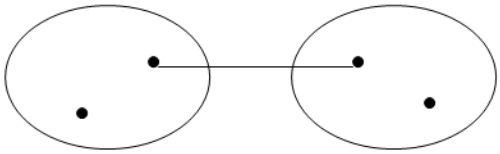
1.1 Types of Clustering

Flat clustering
Flat clustering creates a flat set of clusters without any explicit structure that would relate clusters to each other.
Hard clustering
In hard clustering each document is a member of exactly one cluster.
Soft clustering
In soft clustering a document has fractional membership in several clusters.
K-Means Clustering
Select K random docs {s1, s2,…., sK} as seeds.
Until clustering converges or other stopping criterion:
For each doc di:
Assign di to the cluster cj such that dist (xi, sj) is minimal.
(Update the seeds to the centroid of each cluster)
For each cluster cj
$s_j = \mu(c_j)$
Hierarchical Clustering
Build a tree-based hierarchical taxonomy from a set of documents.

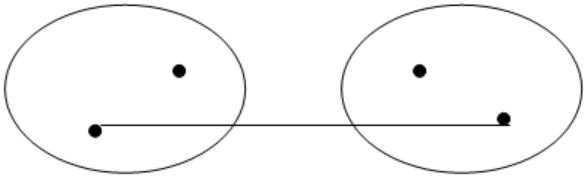


Many variants to defining closest pair of clusters

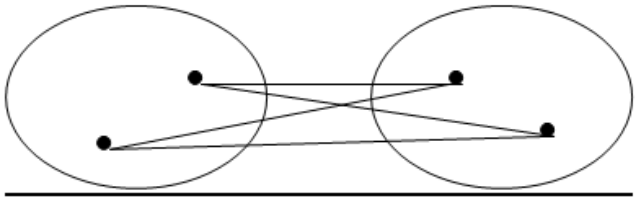
Single-link
In single-link clustering or single-linkage clustering, the similarity of two clusters is the similarity of their most similar members.



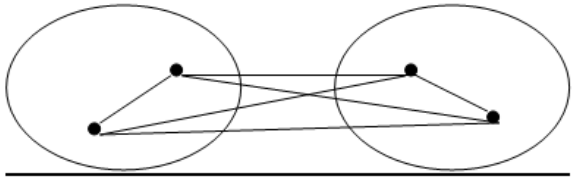
Complete-link
In complete-link clustering or complete-linkage clustering, the similarity of two is the similarity of their most dissimilar members.



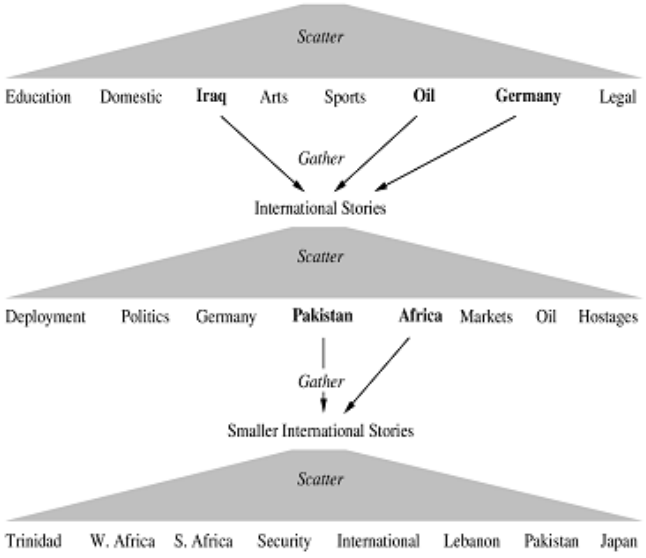
Centroid
average inter-similarity



Group-average
average of all similarities

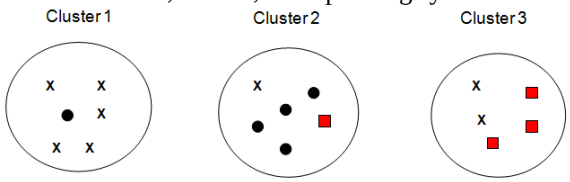


1.2 Application of Clustering
Search result clustering
Search result clustering clusters the search results, so that similar documents appear together.
Scatter-Gather
Clusters the whole collection to get groups of documents that the user can select or gather.



Collection clustering Compute a static hierarchical clustering of a collection that is not influenced by user interactions.

1.3 Evaluation of Clustering
Purity To compute purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by N. For example, the search results for jaguar consists of three classes corresponding to the three senses car, animal, and operating system.



Majority class and number of members of the majority class for the three clusters are:
X, 5 (cluster1) ●, 4 (cluster2) ■, 3 (cluster3)
Purity is $(1/17) \times (5 + 4 + 3) \approx 0.71$
Purity $(\Omega, C) = 1/N \sum_{k} \max_{j} |\omega k \cap cj|$
Where $\Omega = \{\omega 1, \omega 2, \ldots, \omega K\}$ is the set of clusters and $C = \{c1, c2, \ldots, cj\}$ is the set of classes.

2. RELATED WORK ON DOCUMENT CLUSTERING

Framework for Hierarchical Clustering Based Indexing in Search Engines ," Parul Gupta and A.K. Sharma",
Let D={D1, D2,……Dn) be the collection of N textual documents being crawled to which consecutive integers document identifiers 1…n are assigned. Each document Di can be represented by a corresponding set Si such that Si is a set of all the terms contained in Di. Let us denote that set by D* such that D*={S1,S2,…….. Sn}. The similarity of any two documents Si and Sj can be computed using the similarity measure [1]: Similarity measure (Si, Sj) = |Si ^ Sj | / |Si U Sj |
Another work proposed was the reordering algorithm [1] Which partitions the set of documents into k ordered clusters on the basis of similarity measure. According to this algorithm, the biggest document is selected as centroid of the first cluster and n/k1 most similar documents are assigned to this cluster. Then

the biggest document is selected and the same process repeats. The process keeps on repeating until all the k clusters are formed and each cluster gets completed with n/k documents. This algorithm is not effective in clustering the most similar documents. The biggest document may not have similarity with any of the documents but still it is taken as the representative of the cluster. This research work has some shortcomings because selecting the documents from the group and assign a cluster may have the possibility that document belonging to one cluster also have chance to belong to another cluster also.

Keyword-based Document Clustering[23]
In this paper, a new clustering method is proposed that is based on the keyword weighting approach. The clustering algorithm starts from the seed documents and the cluster is expanded by the keyword relationship. The evolution of the cluster stops when no more documents are added to the cluster and irrelevant documents are removed from the cluster candidates.

Keyword-based Weighting Scheme
It is common that terms and their weight values represent a document and <term, weight> pairs are the unique elements of the document vector. When we construct a document vector, term frequency and document frequency are the most important features to calculate the weight of a term. As for the terms and their weight values, the weight value of a term means a ranking score just as an importance factor to the document. So, the term weighting can be seen as an evaluation of the term as a keyword or a stop word to the document. The weighting function
w (t) from a term to its weight is described in expression (1).
w: term → weight (1)
w (t) = 0, if t is a stop word 1, if t is a keyword a, otherwise $0 \leq a \leq 1$

Keyword-based Document Clustering
Keyword-based document clustering creates a cluster by the keywords of each document. Suppose that C is a set of clusters that is finally created by the clustering algorithm. If n is the number of clusters in C, then C is a set of clusters,, …, C1 C_2 C_n
$C = \{C_1, C_{2,\ldots,} C_n\}$
Each cluster is initialized by document that is not assigned to the existing clusters, and is a seed document. When a new cluster is created, expansion and reduction steps are repeated until it reaches a stable state from the start state. In each evolution steps for cluster, is the j-th state of Ci. The characteristic vector of a cluster is a set of <keyword, weight value> pairs that represents the cluster. If is a keyword set of a document and I is a keyword set of cluster, then i is the j-th state of cluster. The disadvantage with this method is that it not suggests any efficient method to calculate the weight of the term because some terms are not so important and contain no meaning to the document therefore assigning a weight to these terms is not a good idea.

5. CONCLUSION AND FUTURE SCOPE

Document clustering (or Text clustering) is automatic document organization, topic extraction and fast information retrieval or filtering. A web search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories, as is achieved by Enterprise Search engines. This paper describes various document clustering techniques and also the working behind the technique so, that programmer can use efficient technique according to his need. There also mentioned a purity check for the cluster created by the clustering technique.

6. REFERENCES

[1]. Fabrizio Silvestri, Raffaele Perego and Salvatore Orlando. "Assigning Document Identifiers to Enhance Compressibility of Web Search Engines Indexes" In the proceedings of SAC, 2004.

[2] .Van Rijsbergen C.J. "Information Retrieval" Butterworth 1979

[3]. Oren Zamir and Oren Etzioni. "Web Document Clustering: A feasibility demonstration" In the proceedings of SIGIR, 1998.

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

[5]. Sanjiv K. Bhatia. "Adaptive KMeans Clustering" American Association for Artificial Intelligence, 2004.

[6]. Bhatia, S.K. and Deougan , J.S. 1998. "Conceptual Clustering in Information Retrieval" IEEE Transactions on Systems, Man and Cybernetics.

[7]. Dan Bladford and Guy Blelloch. "Index compression through document reordering" In IEEE, editor, Proc. Of DCC'02. IEEE, 2002.

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

[9]. Khaled M. Hammouda, Mohamed S. Kamel: Efficient Phrase-Based Document Indexing for Web Document Clustering. IEEE Trans. Knowl. Data Eng. (TKDE) 16(10):1279-1296 (2004).

[10]. Benjamin Fung, Ke Wang, Martin Ester, "Hierarchical Document Clustering Using Frequent Itemsets", May 1, 2003

[11]. Parul Gupta and A.K. Sharma,"A Framework for Hierarchical Clustering Based Indexing in Search Engines", BVICAM's International Journal of Information Technology (BIJIT) June 2012.