제21회 한국정보처리학회 춘계학술발표대회 논문집 제11권 제1호 (2004. 5)

계층적 클러스터링에서 분류 계층 깊이에 관한 연구*

김해남, 이신원, 안동언, 정성종
전북대학교 컴퓨터공학과
e-mail : hnkim@duan.chonbuk.ac.kr

A Study on Cluster Hierarchy Depth in Hierarchical Clustering

Hai-Nan Jin, Shin-won Lee , Dong-Un An，Sung-Jong Chung
Dept. of Computer Engineering ,Chonbuk National University

ABSTRACT

Fast and high-quality document clustering algorithms play an important role in providing data exploration by organizing large amounts of information into a small number of meaningful clusters. In particular, hierarchical clustering provide a view of the data at different levels, making the large document collections are adapted to people's instinctive and interested requires.

Many papers have shown that the hierarchical clustering method takes good-performance, but is limited because of its quadratic time complexity. In contrast, K-means has a time complexity that is linear in the number of documents, but is thought to produce inferior clusters.

Think of the factor of simpleness, high-quality and high-efficiency, we combine the two approaches providing a new system named CONDOR system [10] with hierarchical structure based on document clustering using K-means algorithm to "get the best of both worlds". The performance of CONDOR system is compared with the VIVISIMO hierarchical clustering system [9], and performance is analyzed on feature words selection of specific topics and the optimum hierarchy depth.

1. INRTODUCTION

Document clustering is the operation of generating grouping together related (or similar) documents to generate a category structure. It seeks to minimize within-group variance and maximize between-group variance. It iterates in [1].

There are many clustering method available, which base on different theoretical or empiricism. For given clustering method, there are provided a choice of clustering algorithm. The choice of clustering method will determine the outcome and the choice of algorithm will determine the efficiency.

Section 2 presents a short analysis of the clustering method. Section 3 introduces the CONDOR system's clustering algorithm. Section 4 describes the associated evaluation strategy, shows the comparative clustering result. At last give the conclusion and further work.

2. SURVEY OF CLUSTERING METHODS

Clustering methods are usually categorized according to the type of cluster structure they produce. The simple nonhierarchical methods divide the document collection of N objects into M clusters without overlap. A priori decisions about the number of clusters, cluster size, criterion for cluster membership, and form of cluster representation are required. These factors lead to different clustering results. The computational requirement is $O(NM)$.The example work on the SMART project, described by [7].

The last decade of work on clustering in IR retrieval has concentrated on the hierarchical clustering methods [8]. It has been considered as an improved method with nonparametric. Large CPU time and high memory are required. Hierarchical techniques produce a nested sequence of partitions. There are two main approaches: agglomerative and divisive.

* 본 연구는 한국과학재단 목적기초연구 R01-2003-000-11588-0 지원으로 수행되었음.

There are several methods for determining the distances between clusters [5]. The most common metrics can be divided in two general classes:

- **Graph Methods:** these methods determine intercluster distances using the graph of points in the two clusters. It includes *Single Link, Average Link and Complete Link.*
- **Geometric Methods:** these methods define a cluster center for each cluster and use these cluster centers to determine the distances between clusters. It includes *Centroid, Median and Minimum Variance.*

Useful clustering metrics can usually be described using the Lance-Williams updating formula [2]. The distance d between the new cluster $C_{i,j}$ and any existing cluster C_k is given by:

$$d_{C_{i,j}C_k} = a(i) \bullet d_{C_iC_k} + a(j) \bullet d_{C_jC_k} + b \bullet d_{C_iC_j} + c \,|\, d_{C_iC_k} - d_{C_jC_k} \,|$$

Most of the works on hierarchical clustering have been aimed to develop faster and more efficient algorithms to implement it. In the paper [6] propose a *relative dissimilarity measure* that works better on well-represented categories or groups (including overlapping) than the classic methods.

3. CONDOR SYSTEM CLUSTER ALGORITHM

3.1 System Diagram

CONDOR system uses hierarchical clustering technique to index and retrieval large documents collection. Not only include indexing, query processing and summarization etc., but also achieve interaction by API. Figure 1 shows the main process of CONDOR system.

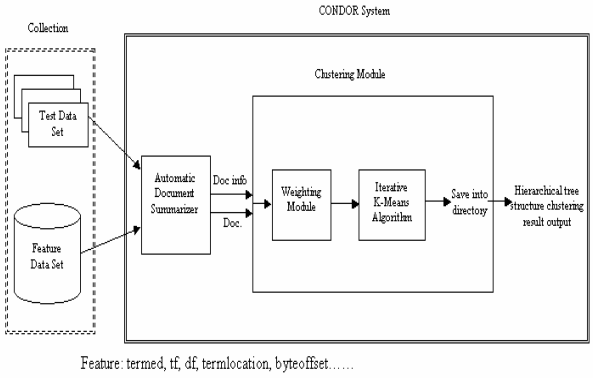


Figure 1: The clustering kernel of CONDOR system

3.2 Documents Clustering with K-means Algorithm

CONDOR system uses the nonhierarchical K-means algorithm to clustering lager practicability web data set to reduce CPU time and computational complexity. It is worthy even though some precision expense.

K-means algorithm is a partition technique. It is based on the idea that a center point can represent a cluster. For K-means we use the notion of a centroid, which is the mean or median point of a group of points. Note that a centroid almost never corresponds to an actual data point. In our system the centroid vector $\vec{c}_j$ is obtained by averaging the weights of the various terms present in the documents of corresponding cluster.

K quantitative clusters are obtained by the first clustering operation, nested re-clustering the sparser clusters that the distance between any pair of documents in the cluster oversteps a threshold. Repeat this process, until the distance value limits in a boundary. The repeat times equals clustering depths. Figure 2 describes this case visually:

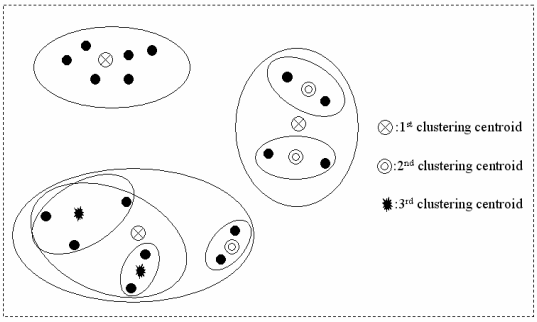


Figure 2: hierarchical clustering

CONDOR system's K-means clustering algorithm is shown in following table:

1) Select number of the cluster: K
2) Obtain K initial cluster centroids
3) Calculate the Euclidean distance (dist) between K cluster centroids and each documents
$$dist(\vec{d}_j, \vec{c}_r) = \sqrt{\sum_{k=1} (d_{kj} - c_{kr})^2}$$
4) Allocation the documents to one of K cluster centroid which has a short distance
$$\underset{\substack{j=1,m \\ r=1,k}}{\operatorname{argmin}} dist(\vec{d}_j, \vec{c}_r)$$
$$d_j \in G_c, \; if \; dist(\vec{d}_j, \vec{c}_r) < dist(\vec{d}_j, \vec{c}_l)$$
$$(for \; all \; l = 1,2,...,k \;\; l \neq r)$$
5) Recalculate the K cluster centroids
$$\vec{c}_r = \frac{1}{|c_r|} \sum_{l=1}^{|k|} \vec{d}_l$$
6) If the distance between old centroids and new centroids is more than θ, go to step 3, else finish the algorithm $if \; \max\delta(c_r^{old}, c_r^{new}) < \theta \; then \; return \; else \; goto 3$
7) For any cluster, if the distance between pairwise documents oversteps boundary then nested reiteratively clustering
8) Save the clustering result with directory

3.3 Output Hierarchical Structure of Clusters

We can see for individual document collection, the clustering depth is different. Experiments show the optimal depth's bound is 3.

We use a cursor to browse the clustering directory. Assign original cluster for cursor, test whether has child traverse uniform depth. According to the cursor's browsing situation, output the tree configuration to achieve alike clustering structure with hierarchical clustering method.

Figure 3 describes the situation with maximum depth is 3

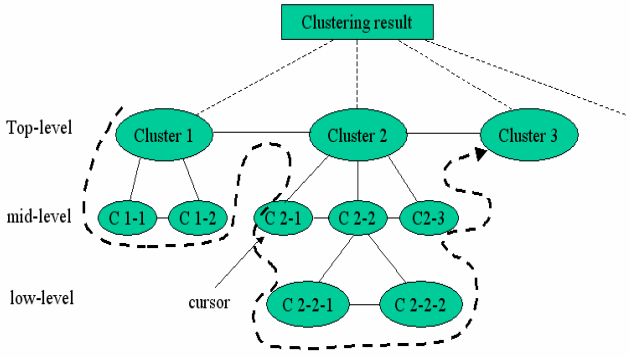


Figure 3: tree configuration

The algorithm can summarize with following table:

```
<Skeleton of application program for accessing hierarchical clustering>
access_current (cid)
{
      get_fist_cluster_this level (cursor, cid)
      if (first_child_cid = have_child (cursor, cid))
        access_current (first_child_cid)
      while (next exist)
      {
            get_next_cluster_this_level (cursor, cid)
            if (first_child_cid = have_child (cursor, cid)
              access_current (first_child_cid)
      }
}
```

4. Performance

In this section we discuss different performance issues, and then perform experiments using the suggested parameter settings. For individual document collection, analyze the clustering result in dynamic dimensional space. (Unfixed index items are obtained by lexical analysis with a weight boundary.) We used the news data is reported in some Korean newspapers from Mar.2002 to Sept.2003 to compare tree configuration document clustering by K-means algorithm with the classical hierarchical clustering. We invited some professionals and some tenderfoots using manual method to evaluate CONDOR system's performance, examined the feasibility of the hierarchical structure based on document clustering using K-means method.

4.1 Performance Issues

- Optimal Feature Words: VIVISIMO system is an automatic hierarchical clustering technique, uses less than 2 keywords to character individual cluster. The performance is very good for the English query, but for Korean need adjective, verb etc. besides noun. In order to improve CONDOR performance, extract the nominal feature words by a thesaurus. We set maximum number is 3. The index word's weight is computed with

$$weight = \frac{tf}{tf+2} \times \frac{df+2}{df}$$

To reduce the influence of the local weight on the document weight, we set *term frequency* amounts $tf/(tf+2)$. Consequently improve the essentiality of the global weight of the document's weight df. Similar information iterates in [3].

So the feature words selection algorithm is summarized with the following table:

1) Output noun using thesaurus

$if\ (term_{i-C_j} \in noun)\ then\ (term_i \rightarrow TermS_{C_j})$

$TermS_{C_j}$: *jth cluster's index words collection*

$term_{i-C_j}$: *ith index word of jth cluster*

i : *the number of special cluster's index words*

j : *the number of clusters*

2) Remove the words are used by superstratum directory

$if\ (TermS_{C_{j+1}} \notin TermS_{C_j})\ then\ TermS_{C_{j+1}}$

3) Select 3 words with the highest weight as the cluster feature words

Experiment show that CONDOR's performance using thesaurus is better than VIVISIMO in extracting the feature words of special cluster.

- Optimum initial K value and Hierarchy Depth: This paper presented CONDOR evaluation base on setting optimum initial K value and clustering depth mainly

4.2 Experiment and Evaluation

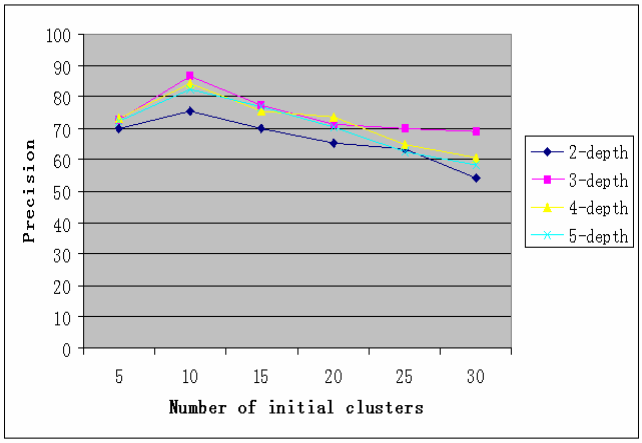
We use large number of homonym, nomenclature of local subject, synonym and free stochastic words to evaluate clustering precision. The following table gives out comparative examination result base on four partial popular Korean queries.

Query	Cluster number	2-15	3-15	Topic
유산	Total cluster numbers	32	33	Bequest, Maternal inheritance, Culture tradition, Ashes..
	Correct cluster numbers	20	23	
장수	Total cluster numbers	28	33	Company name, Health, Area, Logevity..
	Correct cluster numbers	13	24	
조선	Total cluster numbers	30	31	Nation name, Newspaper name, Shipbuilding..
	Correct cluster numbers	26	26	
화장	Total cluster numbers	32	35	Cosmetic, Korean tradition dress, Cremation..
	Correct cluster numbers	27	29	

We used the labeled initial centroids number of the clusters to present clustering precision, as a comparative measure bases on different clustering depths. Following table shows these results.

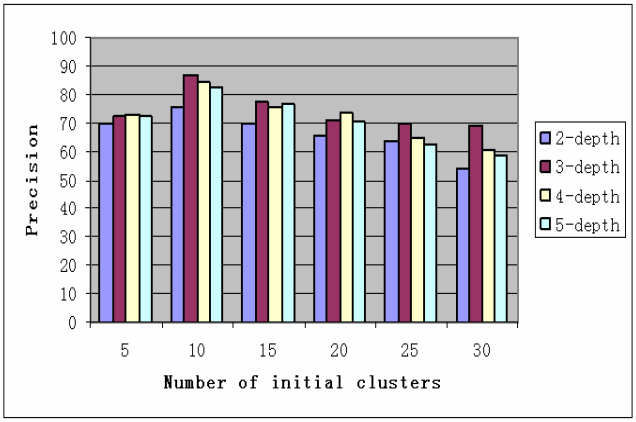
Number of initial clusters	Precision (%)			
	2-depth	3-depth	4-depth	5-depth
5	69.73	72.54	73.12	72.14
10	75.42	86.74	87.04	82.48
15	70.00	77.29	75.24	76.68
20	65.32	71.38	73.38	70.49
25	63.49	70.12	64.67	62.38
30	54.21	68.89	60.43	58.49

Experiment shows the optimal performance of CONDOR system is obtained when sets 10 initial centroids. It indicates larger K value doesn't mark better clustering effect.

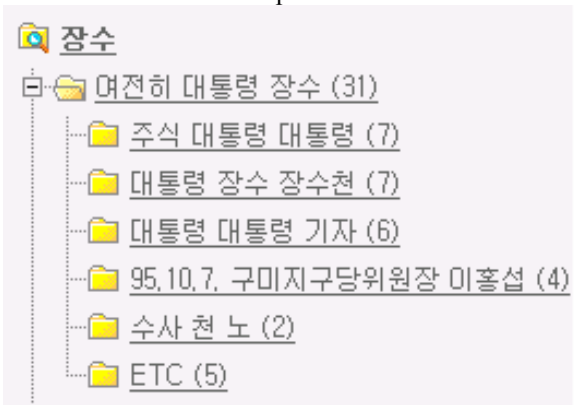


Experiment shows 3-depth clustering result like 4-depth

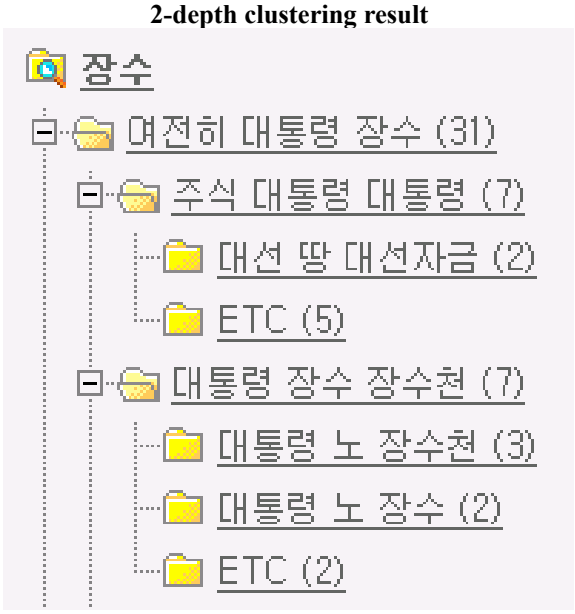
and 5-depth clustering result. So we can infer CONDOR system optimal depth is 3.



Following two figures show the clustering effect for 10 initial centroids and 2 or 3 depth.



2-depth clustering result



3-depth clustering result

Experiment shows CONDOR system's precision overruns 80%. We believe more high-quality performance can obtained by properly tuning CONDOR system's parameters.

5. CONCLUSIONS AND FURTHER WORK

We have presented a new idea based on hierarchical documents clustering using K-means algorithm. Complexity analysis and experiments show this method is feasible, not only satisfy human instinctive and conventional information retrieval require but also reduce expending in time and memory which using more complex hierarchical clustering method though the performance of CONDOR system is less under the VIVISIMO system.

Though great improvement, not achieve optimal clustering result. How to combine K-means algorithm's high efficiency and hierarchical clustering method's high performance is still a hard work. From the evaluated result, we can conclude that the choices of initial centroids and optimum K value, improve document weight computation metric and evaluation method, using the improved high-powered clustering algorithm (such as bisecting K-means mentioned in [4]), the thesaurus redaction and the selection strategy of feature words of special topic can affect the performance of CONDOR system's performance and efficiency. The hierarchical structure based on document clustering using K-means algorithm is only a transition from simple K-mean to complicated hierarchical clustering method. It is hoped that future work will lead to an effective operation based on this idea that can then be validated on large and fast web data collection till resolve the problem of hierarchical method's efficiency.

REFERENCE

[1] Andergerg, M.R. "Cluster analysis for applications" New York Academic,1973

[2] G. N.Lance and W.T.Williams. "A general theory of classificatory sorting strategies". 1: Hierarchical systems. Computer Journal, 9:373-380, 1967

[3] Ji Hyun Go, "A Study on the Index Terms Weighting Scheme Using Latent Semantic Indexing Approach on the Document Clustering", 정보처리 2003 학회논문지 B (인공지능), Dept. of Information Communications Engineering and Computer Engineering, Chonbuk National University.

[4] Michael Steinbach, George Karypis, Vipin Kumar, "A Comparison of Document Clustering Techniques", Technical Report #00_034. Department of Computer Science and Engineering, University of Minnesota, 2000.

[5] Ramon A. Mollineda, Enrique Vidal. "A relative approach to hierarchical clustering", 2000.

[6] Willett, p. 1988. "Recent Trends in Hierarchic Document Clustering: A Critical Review." information processing & Management, 24 (5), 577-97

[7] Salton, G., ed. 1971. "The SMART Retrieval System." Englewood Cliffs, N.J.: Prentice Hall.

[8] Murtagh F., "Multidimensional Clustering Algorithms", Physica-verlag, 1985

[9] Vivisimo http://vivisimo.com

[10] S.C Park, D. U. An, "Conodor Information Retrieval System", Journal of the Korea Society Industrial Information Systems, Vo1.8, No.4, pp 31~37, 2003