Selection of Cluster Topic Words in Hierarchical Clustering using K-Means Algorithm

Shin-won Lee, Sang-seon Yi, Dong-un An and Sung-jong Chung
Dept of Computer Engineering, Chonbuk National University, Jeonju, 561-756, Korea
Tel : +82-063-270-2416 Fax : +82-063-270-2394 E-mail: swlee@duan.chonbuk.ac.kr

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. Hierarchical clustering improves the performance of retrieval and makes that users can understand easily. For outperforming of clustering, we implemented hierarchical structure with variety and readability, by careful selection of cluster topic words and deciding the number of clusters dynamically. It is important to select topic words because hierarchical clustering structure is summarizes result of searching. We made choice of noun word as a cluster topic word. The quality of topic words is incresed 33% as follows. As the topic word of each cluster, the only noun word is extracted for the top-level cluster and the used topic words for the children clusters were not reused.

Keyword: Hierarchical Clustering, Cluster Topic Words

1. INTRODUCTION

If user inputs query then the common information retrieval system outputs long search result. So user can find proper document. We need times and efforts, searching and organizing for many topic of related document. So we need the study that system is organized automatically sets of document.

Document clustering is a technique that classifies a set of documents by subject. It is used for analyzing documents structure or improving efficiency of in information retrieval. Hierarchical clustering has better retrieval performance than that of non-hierarchical clustering. Also, users can read easily the result of retrieval because it displays the documents structure hierarchically.

Since the work to do clustering correctly lots of documents takes a long time, it is hard to implement a real time system to show the retrieved result such as WEB Search Engine. However, for improving the organization quality structurally and keeping the retrieval performance high, we need fast and correct clustering algorithm. So we selected K-Means algorithms.

Since the only few words represent the content of cluster implicitly, we need to consider the selecting of cluster topic word. In this paper, the excellent quality of topic words should be made as follows. The only noun word is extracted for selecting each cluster topic word by using dictionary and the used words in the parent clusters did not reused for the children clusters.

This paper is organized as follows. Chapter 2 presents a short analysis of the clustering method. Chapter 3 introduces architecture, clustering module and cluster topic selection of the Condor system. Chapter 4 describes the associated evaluation strategy, shows the comparative clustering result. Finally Chapter 5 concludes.

2. RELATED WORK

There are two approaches of analyzing clusters. The non-hierarchical method divides the data set of N objects into M clusters. The hierarchical method produces a nested data set in which pairs of items or clusters are consecutively linked until every item in data set is connected.

Hierarchical clustering starts with each document by building one cluster and repeat the process of merging two clusters which have high similarity until one cluster remains. There are several methods of hierarchical clustering. According to the criteria of selecting two clusters which have high similarity, they are called 'single link', 'complete link', 'group average link', and 'Ward's method'.[2]

Non-hierarchical clustering starts with randomly selected initial clusters and repeats the process of relocating clusters. There are several methods of non-hierarchical clustering such as 'single pass method', 'K-Means Algorithm'.[3] This method is often used for clustering in real time such as fast processing of lots of document.[4][5][6]

Vivisimo[1] provides meta search capability and hierarchical clustering functionality of documents automatically in real-time. This method selects one more word as cluster topic word and shows clustering quality. However, in case of Korean language, it selects other a part of speech as well as noun, adverb and adjective for cluster topic word, so that it doesn't show the good clustering quality than that of in English.

3. CONDOR SYSTEM

We experimented the proposed approach using Condor Information Search Engine which is developed by the laboratory of Intelligence Engineering of Chonbuk National University, The Language Technology Institute of Carnegie Mellon University, USA, and the SearchLine Inc.[7][8][9] Fig 1 is result of Condor system.

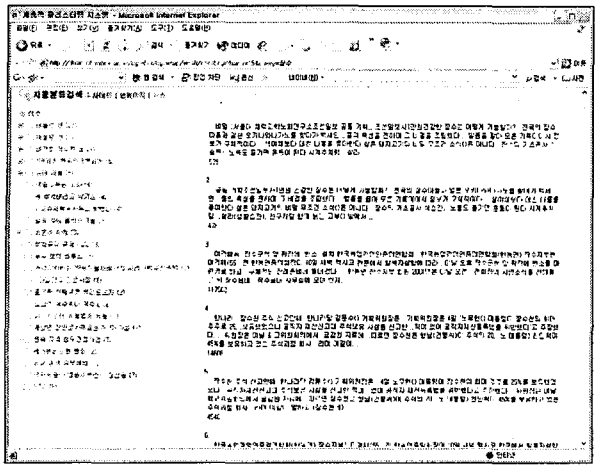

Fig 1 Condor System

In Fig 2, The Condor Search Engine is composed of Index Module, Search Engine Module, and User Interface. The Hierarchical Clustering Parts of Condor Search Engine are composed of indexing module, query parser, summary generation component and API.

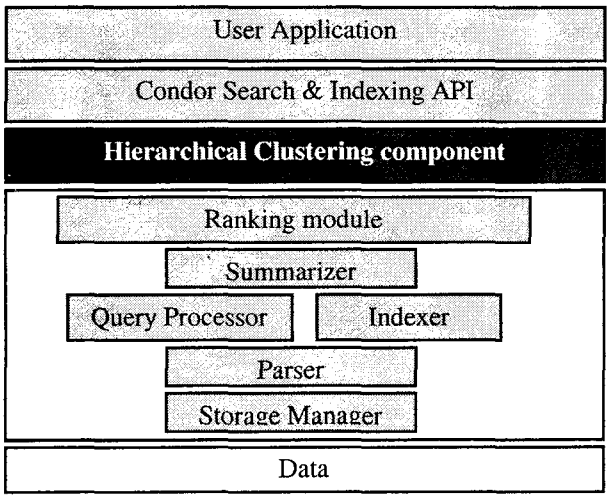

Fig 2 Condor System Architecture

The Condor system makes use of K-Means algorithm to construct a hierarchical structure. Comparing with other hierarchical clustering algorithm, the K-Means algorithm has strong points that it is easy to implement, is speedy to process and is capable of do clustering work on real-time for the sake of relatively low precision.

For improving the clustering performance, when the hierarchical cluster structure is implemented, we excluded the search keyword as the cluster topic word. Because the target documents that would be clustered is the retrieved document of search keyword, it is more effective that clusters the documents by excluding the search keyword.

For the improvement of clustering, in this paper, we clustered documents hierarchically and show the cluster topics as noun words except the search keyword. According to the retrieved documents, the decision of initial number of clusters dynamically made better results than the fixed number of initial clusters. K-means clustering algorithm of Condor system is shown in the following table.

1 Select number of initial cluster. If document number of result sets is n, number of initial cluster is experimented each $\log_e n, \sqrt{n}, \frac{n}{40}$.

2 Calculate the Euclidean distance between K cluster centroid and each document.

$$dist(D_i, C_j) = \frac{\sum_{k=1}^{n}(D_{i,k} \times C_{j,k})}{\sum_{k=1}^{n}(D_i,k)^2 \times \sum_{k=1}^{n}(C_j,k)^2}$$

i : 1, 2, ..., n n : total document number
j : 1, 2, ..., K k : centroid number

3 Allocation the documents to one of K cluster centroid which has a short distance.

$$\arg\min_{\substack{i=1,n \\ j=1,k}} dist(\vec{d_i}, \vec{c_j})$$

$$d_i \in \underset{i=1,n}{G_{c_j}} \quad if\ dist(d_i, c_j) < dist(d_i, c_l)$$

$$for\ all\ l = 1,2,\cdots,k \quad l \neq j$$

4 Recalculate the K cluster centroid.

$$\vec{C}_j = \frac{1}{|C_j|}\sum_{l=1}^{|C_i|}\vec{d}_l$$

5 If the distance between old centroids and new centroids is more than , go to step 3, else finish the algorithm

$$if\ \max\delta(\overline{C}_j^{old}, \overline{C}_j^{new}) < \theta\ \ then\ return$$

$$else\ goto\ \text{¨é}$$

6 If similarity of document in cluster is less than threshold, cluster again in cluster

7 Save the clustering result with tree.

Total cluster depth is maximum 3 by similarity conditions of document set. Some cluster node is classify 2, 3 depths or 1 depth depending on the documents status. Fig 3 presents hierarchical clustering using K-Means algorithms.

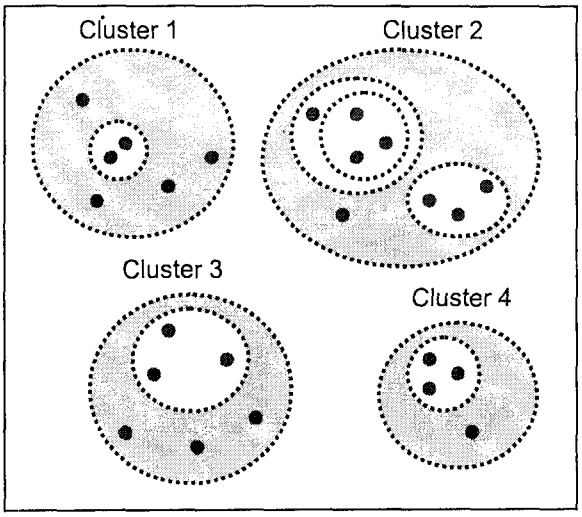

Fig 3 Hierarchical Clustering

After classifying the documents into clusters, we select topic words of each cluster. It is important to select topic words because hierarchical clustering structure is summarizes search result. We selected 3 words that have high weight in each cluster topic. The word's weight is computed with the following equation (1).

$$weight = \frac{tf}{tf + 2.0} \times \frac{df + 2.0}{df} \qquad (1)$$

We made choice of noun word as a cluster topic word. In case of verb or adverb, those words can have ambiguous meaning, so that those are not suitable for summarizing the cluster. In child clusters, the used words in parent clusters need not reused for outperforming quality of topic word because overlapping words are hard to divide.

1 Extract noun words using dictionary

$if\,(term_{j/C_i} \in Noun)\ then\ Terms_{C_i}$

$Terms_{C_i} : i_{st}$ *Term collection in Cluster*$_i$

$term_{j/C_i} : j_{st}$ *Term collection in Cluster*$_i$

i : *number of Cluster*

j : *number of words in Cluster*

2 In child clusters, the used words in parent clusters need not reused.

$if\,(Terms_{C_{i+1}} \notin Terms_{C_i})\ then\ Terms_{C_{i+1}}$

3 Select 3 topic words in weight order.

4. EXPERIMENTS

We made use of the Korean Newspapers (Cho Sun, Joong-Ang, Dong-A Daily, Hankeyre, HanKook, MunHwa) from May to November in the year of 2003 as experimental data and made experiments comparing with the clustering results that using based on introducing noun extraction skill and with/without topic words duplication removal mechanism, respectively.

4.1 Clustering Performance Experiments

We limited the maximum number of retrieval document set (n) as 700. We experimented the discovering the best clustering quality by changing the K clusters depending on the number of retrieved documents. For measuring efficiency of clustering, we used homonym as search keyword and found the results that are quite different in each cluster. Table 1 is queries that used experiments. And Table 2 is results.

Table 1 Queries

Query #	Query	Means
Q_1	(Yoo-san)	abortion, property, lactic acid, ...
Q_2	(Jang-soo)	long life, country name, ...
Q_3	(Hwa-jang)	make-up, cremation
Q_4	(Pa-ri)	fly, Paris

Table 2 Clustering Performance Experiments

Query		$k = \log_e n$	$k = \sqrt{n}$	$k = \frac{n}{40}$
Q_1	TCN	34	83	34
	CCN	13	38	19
	Precision	0.38	0.45	0.56
Q_2	TCN	64	88	55
	CCN	32	36	18
	Precision	0.5	0.41	0.33
Q_3	TCN	27	68	27
	CCN	17	27	15
	Precision	0.63	0.40	0.56
Q_4	TCN	41	101	42
	CCN	20	44	19
	Precision	0.49	0.44	0.45
Total Precision		**0.49**	0.43	0.45

(TCN: Total Cluster Number, CCN: Correct Cluster Number)

As the experiments result, the best performance is acquired in case of $k = \log_e n$. The reason is that the characteristics of log function, that is classify documents in detail, when the number of document is few, it can classify documents variously and when the number of documents is big, it limits the number of clusters.

4.2 Cluster Topic Selection

We calculate precision for measuring the best cluster topic words by counting the correct words of cluster topics among the 3 representative words. If topic words are duplicated topic word in parent cluster or those words are not representative words in the meaning, the words belong to the wrong cluster topic words.

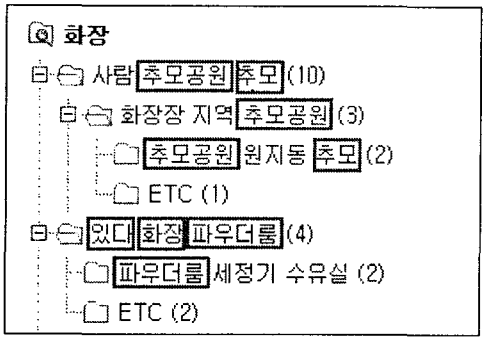

Fig. 4 Original system

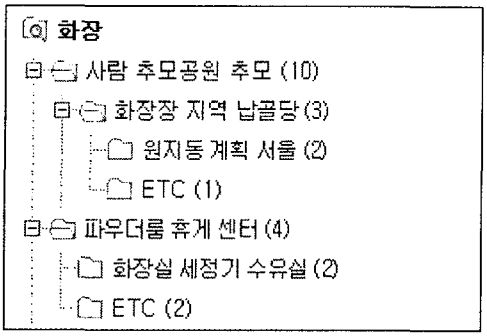

Fig. 5 Applied method system

Table 3 Original Clustering Results

	Q_1	Q_2	Q_3	Q_4	Total
TTN	89	114	90	90	383
CTN	51	63	34	39	187
Precision	0.57	0.55	0.38	0.43	0.49

(TTN: Total Topic Number, CTN: Correct Cluster Number)

Table 4 Extracted Noun Words

	Q_1	Q_2	Q_3	Q_4	Total
TTN	95	105	89	85	374
CTN	50	69	53	57	229
Precision	0.52	0.66	0.60	0.67	0.61

Table 5 Removal of representative word of cluster topic

	Q_1	Q_2	Q_3	Q_4	Total
TTN	84	94	76	73	327
CTN	50	70	51	60	231
Precision	0.60	0.74	0.67	0.82	0.70

Table 6 Applied both methods

	Q_1	Q_2	Q_3	Q_4	Total
TTN	95	108	89	85	377
CTN	61	81	59	72	273
Precision	0.64	0.75	0.66	0.84	0.72

Fig 4 and 5 are capture of system. And Table 3, 4, 5 and 6 are results. The experiment results showed that the system that noun word extraction and removal of representative word of cluster topic has 33% higher quality than the opposite because we remove unnecessary and repetitive term. The case of selection only noun words improves performance because verb or adverb can have ambiguous meanings, which are not suitable for summarizing the cluster. The duplicative topic word removal method is better performance for showing much information and classifying clusters more effectively. The evaluation is done by human beings.

5. CONCLUSIONS

This paper proposes a clustering method to enhance performance. It shows clustering result except query and decides that the number of initial clusters is proportional to the number of documents. Also, when we select topic word of cluster, it extracts only noun. And in child clusters, the used words in parent clusters need not reused

Acknowledgements

This work was supported by grant No. R01-2003-000-11588-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

References

[1] Vivisimo http://www.vivisimo.com/

[2] Qin He, "A Review of Clustering Algorithms as Applied in IR," UIUCLIS-1999/6+IRG

[3] Tapas Kanung, "The Analysis of a Simple k-Means Clustering Algorithms" in Proceedings of the sixteenth annual symposium on Computational geometry, 2000

[4] Baeza-Yates, Rebeiro-Neto, "Modern Information Retrieval", Addison-Wesley

[5] Hyung Jin Oh "Analysis of Document Clustering Varing Cluster Centroid Decisions", Proceedings of IEEK Summer Conference 2002

[6] khaled Alsabti, 1998, Sanjay Ranka, Vineet Singh, An Efficient K-Means Clustering Algorithm, IIPS 11th International Parallel Processing Symposium.

[7] Soon Cheol Park, Dong-un An, "Concor Information Retrieval System", Korea Society Industrial Information Systems. Vol 8 No.4, 2003.

[8] Sang-seon Yi, Shin-won Lee, Dong-un An, Surg-jong Chung, "A Study on Cluster Topic Selection in Hierarchical Clustering", Proceedings of the 20th KIPS Spring Conference

[9] Hai-nan Jin, Shin-won Lee, Dong-un An, Surg-jong Chung, "A Study on Cluster Hierarchy Depth in Hierarchical Clustering", Proceedings of the 20th KIPS Spring Conference