논문 2008-45SD-2-21

K-means 알고리즘을 이용한 계층적 클러스터링에서의 클러스터 계층 깊이 선택

(Selection of Cluster Hierarchy Depth in Hierarchical Clustering using K-Means Algorithm)

이 원 휘*, 이 신 원**, 정 성 종**, 안 동 언**

(WonHee Lee, ShinWon Lee, SungJong Chung, and DongUn An)

요 약

정보통신의 기술이 발달하면서 정보의 양이 많아지고 사용자의 질의에 대한 검색 결과 리스트도 많이 추출되므로 빠르고 고품질의 문서 클러스터링 알고리즘이 중요한 역할을 하고 있다. 많은 논문들이 계층적 클러스터링 방법을 이용하여 좋은 성능을 보이지만 시간이 많이 소요된다. 반면 K-means 알고리즘은 시간 복잡도를 줄일 수 있는 방법이다. 본 논문에서는 계층적 클러스터링 시스템인 콘도르(Condor) 시스템에서 K-Means 알고리즘을 이용하여 효율적으로 정보 검색을 하고 검색결과를 계층적으로 볼 수 있도록 구현하였다. 이 시스템은 K-Means Algorithm을 이용하였으며 클러스터 계층 깊이와 초기값을 조절하여 더 나은 성능을 보임을 알 수 있다.

Abstract

Many papers have shown that the hierarchical clustering method takes good-performance, but is limited because of its quadratic time complexity. In contrast, with a large number of variables, K-means reduces a time complexity. Think of the factor of simplify, high-quality and high-efficiency, we combine the two approaches providing a new system named CONDOR system with hierarchical structure based on document clustering using K-means algorithm. Evaluated the performance on different hierarchy depth and initial uncertain centroid number based on variational relative document amount correspond to given queries. Comparing with regular method that the initial centroids have been established in advance, our method performance has been improved a lot.

Keywords: K-means, Document clustering, Information Retrieval, Hierarchical clustering

I. Introduction

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.

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. We have put forward a kind of view informer research that was to construct a hierarchical structure based on non-hierarchical K-means clustering algorithm.

Section 2 presents a short analysis of the clustering method. Section 3 introduces architecture, clustering module of the CONDOR system. Section 4 describes the associated evaluation strategy, shows

* 학생회원, ** 정회원, 전북대학교 전자정보공학부
(Dept. of Electronics & Information Engineering, Chonbuk National University)
접수일자: 2007년11월13일, 수정완료일 : 2008년2월5일

the comparative clustering result. Finally Section 5 concludes.

Ⅱ. Survey of Clustering Methods

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'[5~7].

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'[10]. This method is often used for clustering in real time such as fast processing of lots of document[1, 3~4].

Vivisimo[11] 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.

Ⅲ. CONDOR System and Cluster Algorithm

1. System Diagram

We experiment with the proposed approach using CONDOR Information Search Engine which is developed by the laboratory of Intelligence Engineering of Chonbuk National University, The

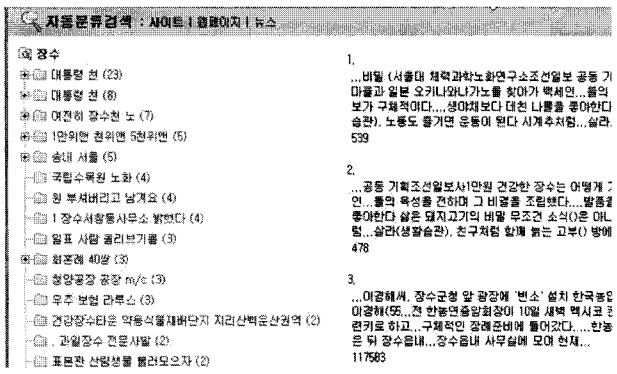


그림 1. 콘도르 시스템의 실행 결과
Fig. 1. Result of CONDOR System.

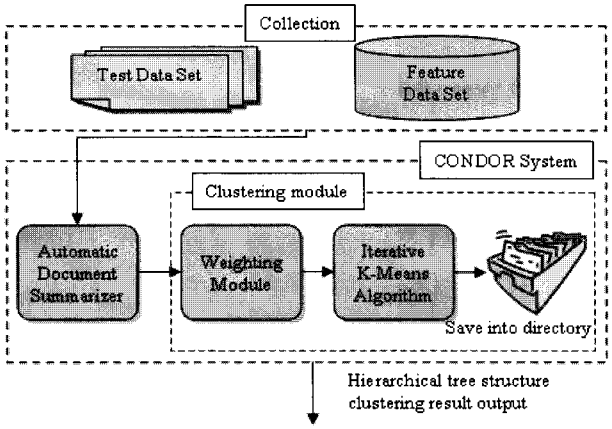


그림 2. 콘도르 시스템 클러스터링 커널
Fig. 2. Clustering kernel of CONDOR.

Language Technology Institute of Carnegie Mellon University, USA, and the SearchLine Inc[2, 8~9]. Figure 1 is result of CONDOR system.

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

The CONDOR system makes use of K-Means algorithm to construct a hierarchical structure. This 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 2 shows the clustering kernel of CONDOR system.

2. Documents Clustering with K-means Algorithm

CONDOR system uses the non-hierarchical K-means algorithm to clustering larger practicability

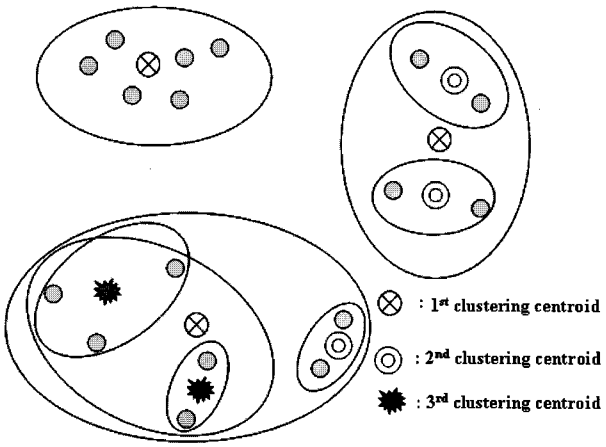


그림 3. 클러스터링 중심 재생성
Fig. 3. re-clustering of Clustering centroid.

표 1. 콘도르 시스템의 K-Means 알고리즘
Table 1. K-Means algorithm of CONDOR.

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

$$dist(\vec{d}_j,\vec{c}_r)=\sqrt{\sum_{k=1}(d_{k,j}-c_{k,r})^2}$$

4. Allocation the documents to one of K cluster centroids which has a short distance

$$\arg\min dist(\vec{d_j},\vec{c_r})$$
$$j=1,m$$
$$r=1,k$$
$$d_j\in G_c, if\ dist(\vec{d_j},\vec{c_r})<dist(\vec{d_j},\vec{c_l})$$
$$(for\ all\ l=1,2,...,k\ l\neq r)$$

5. Recalculate the K cluster centroids

$$\vec{c}_r=\frac{1}{|c_r|}\sum_{l=1}^{|c_l|}\vec{d}_l$$

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 go to step 3
7. For any cluster, if the distance between otherwise documents oversteps boundary then nested reiteratively clustering
8. Save the clustering result with directory

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 3 describes this case visually.

CONDOR system's K-means clustering algorithm is shown in following table 1.

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 or not having 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 4 describes the situation with maximum depth is 3.

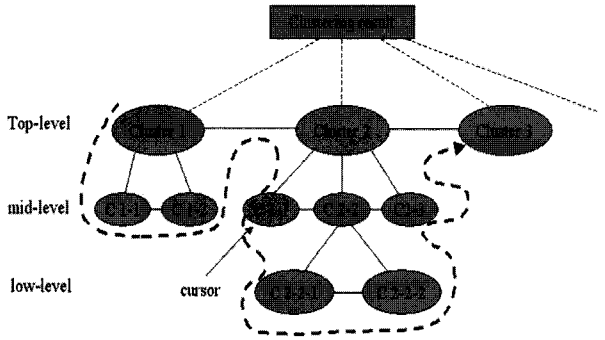


그림 4. 깊이 3에서의 구조
Fig. 4. The situation with maximum depth is 3.

Ⅳ. 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. 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.

1. Performance Issues

We extract 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, we extract the nominal feature words by a thesaurus. We set maximum number is 3. The index word's weight is computed with the following equation (1).

$$weight = \frac{tf}{tf+2} \times \frac{df+2}{df} \quad (1)$$

To reduce the influence of the local weight on the document weight, we set term frequency amounts $tf/(tf+2)$. Consequently we improve the essentiality of the global weight of the document's weight df. Similar information iterates in[9].

So the feature words selection algorithm is summarized with the following table 2.

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

This paper presented CONDOR evaluation base on setting K value and clustering depth mainly. K quantitative initial centroid are obtained by grouping together related documents relative to user's query. K is calculated with 5, 10, 15 expressions corresponding to variable relative documents. One of our major tasks is the choice of the feasible strategy from three above expressions relative to retrieved different documents amount.

CONDOR system uses nested re-partition K-means clustering method, re-clustering the sparse 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. Then the repeat times equals clustering depths.

표 2. 자질 선택 알고리즘
Table 2. Feature word selection algorithm.

1. Output noun using thesaurus

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

$TermS_{c_j}$: j_{th} *cluster's index words collection*

$term_{i-c_j}$: i_{th} *index word of* j_{th} *cluster*

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

j : *the number of clusters*

2. Remove the words are used by superstratum directory

$if\ (Term_{c_{j+1}} \notin TermS_{c_j})\ then\quad TermS_{c_{j+1}}$

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

2. Experiment and Evaluation

We used large number of homonym, nomenclature of local subject, synonym and stochastic words to evaluate clustering precision. Generally, user is interested in the portion of retrieved documents corresponding to user's query that usually relative to anterior range with decreasing similarity order. In order to improve the efficiency, we do not process all retrieved documents. Bite off excrescent document portion, only clustering determinative number of documents. We limited the largest number is equal to 400. Experiment shows 3-depth clustering result is similar to much higher depth clustering result. So we can infer CONDOR system optimal depth is 3. There is no uniform criterion for initial centroid number confirmation, only select appropriate strategy provided above for the range of relative document amount. Experiments show optimal strategy is 5, 10, 15 when clustering document number is limited in 100, 200,

표 3. 깊이 3에서의 평균 정확율(%)
Table 3. 3-depth average precision(%).

document number(n)	number of initial clusters		
	5	10	15
100	82	87.5	84
200	75	79.2	77.8
400	59.1	70.3	64

표 4. 실험결과(초기 크러스터 수 = 10)
Table 4. Examination result.
(number of initial cluster = 10)

Query	Cluster number	Depth 2	Depth 3	Meaning
유산	Total	23	30	Bequest, Maternal inheritance,
	Correct	20	25	Culture tradition, Ashes, etc.
지구	Total	34	48	earth, area, rift, persistence,
	Correct	28	42	etc.
조선	Total	33	41	Nation name, Newspaper
	Correct	22	26	name, Shipbuilding, etc.
화장	Total	29	39	Cosmetic, Korean tradition
	Correct	14	22	dress, Cremation

and 400 respectively. Experiment shows CONDOR system's precision achieves 87.5% based on these techniques. We believe more high-quality performance can obtain by properly tuning parameters.

Complexity analysis and experiments show this method is feasible, not only satisfy human instinctive conventional IR but also reduce expending in time and memory.

The following table gives out comparative examination result base on four partial popular Korean queries.

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.

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

표 5. 클러스터링 깊이
Table 5. Different clustering depths.

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	88.74	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

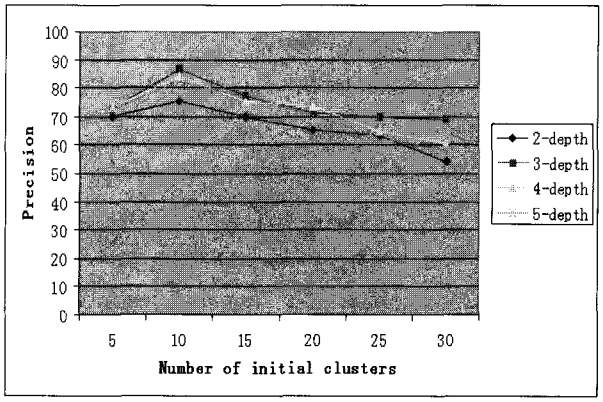


그림 5. 실험 결과
Fig. 5. Experiment result.

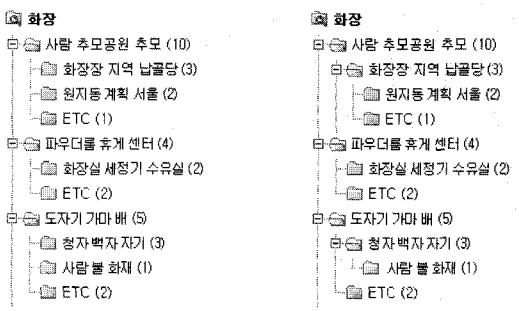


그림 6. 깊이2와 깊이3의 클러스터링 결과
Fig. 6. 2-depth and 3-depth clustering result.

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

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

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

V. Conclusions and Further Work

We have presented a new ideabased 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 educe expending in time and memory which using more complex hierarchical clustering method.

Though great improvement, not achieve optimal clustering. How to combine K-means algorithm's high efficiency and hierarchical clustering method's high performance is still a hard work. The hierarchical structure based on document clustering using K-means algorithm is only a transition from simple K-mean to complicated hierarchical clustering method. Stochastic initial centroid number based on different retrieved documents amount reduces the error of artificial decision. It is hoped that future work will lead to an effective operation based on these ideas that can be validated on large and fast web data collection till resolve the problem of hierarchical method's efficiency.

참 고 문 헌

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

[2] Hai-nan Jin, Shin-won Lee, Dong-un An, Sung-jong Chung, "*A Study on Cluster Hierarchy Depth in Hierarchical Clustering*," Proceedings of the 20th KIPS Spring Conference, 2004.

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

[4] KhaledAlsabti, Sanjay Ranka, Vineet Singh, "*An Efficient K-Means Clustering Algorithm*," IIPS 11th International Parallel Processing Symposium, 1998.

[5] 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.

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

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

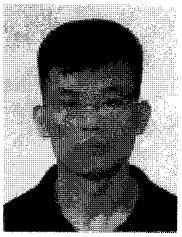
[8] Sang-seon Yi, Shin-won Lee, Dong-un An, Sung-jong Chung, "*A Study on Cluster Topic Selection in Hierarchical Clustering*," Proceedings of the 20th KIPS Spring Conference, 2004.

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

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

[11] Vivisimo http://vivisimo.com

저 자 소 개



이 원 휘(학생회원)

1997년 전주대학교 경영학과 학사 졸업.
1999년 전주대학교 컴퓨터공학과 석사 졸업.
2007년 전북대학교 컴퓨터공학과 박사 수료.

<주관심분야 : 정보검색, 문서분류, 문서요약>

이 신 원(정회원)

1990년 전북대학교 전산통계학과 학사 졸업.
1992년 전북대학교 전산통계학과 석사 졸업.
2005년 전북대학교 컴퓨터공학과 박사 졸업.

<주관심분야 : 정보검색, 한국어정보처리>

안 동 언(정회원)

1981년 한양대학교 전자공학과 학사 졸업.
1987년 KAIST 전산학과 석사 졸업.
1995년 KAIST 전산학과 박사 졸업.

1995년~현재 전북대학교 전자정보공학부 교수
<주관심분야 : 정보검색, 한국어정보처리, 문서분류, 문서요약>

정 성 종(정회원)

1975년 한양대학교 전기공학과 학사 졸업.
1981년 휴스턴대학교 전자공학과 석사 졸업.
1988년 충남대학교 전산공학과 박사 졸업.

1985년~현재 전북대학교 전자정보공학부 교수
<주관심분야 : 정보검색, Grids>