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. INTRODUCTION

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.
There are several methods for determining the distances between clusters [5]. The most common metrics can be divided into 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 $C_{ij}$ and any existing cluster $C_k$ is given by:

$$d_{C_{ij}, C_k} = a(i) \cdot d_{C_{ij}, C_k} + a(j) \cdot d_{C_{ij}, C_k} + b \cdot d_{C_{ij}, C_k} + c \cdot |d_{C_{ij}, C_k} - d_{C_{ij}, C_k}|$$

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.

3.2 Documents Clustering with K-means Algorithm

CONDOR system uses the nonhierarchical K-means algorithm to clustering larger 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 $c_j$ 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.

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.
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 characterize 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

\[
\text{weight} = \frac{tf}{tf + 2} \times \frac{df + 2}{df}
\]

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

   \( \text{if} \ (\text{term} \in \text{thesaurus}) \ \text{then} \ (\text{term} \rightarrow \text{Term}_{i,j}) \)

   \( \text{Term}_{i,j} \) : jth index words collection

   \( \text{term}_{i,j} \) : jth 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 superfluous directive

   \( \text{if} \ (\text{Term}_{i,j} \in \text{Term}_{i,j}) \ \text{then} \ \text{Term}_{i,j} \)

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.

<table>
<thead>
<tr>
<th>Query</th>
<th>Cluster number</th>
<th>2-15</th>
<th>3-15</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>水産</td>
<td>Total cluster numbers</td>
<td>32</td>
<td>33</td>
<td>Bequest, Maternal inheritance, Culture tradition, Atheism..</td>
</tr>
<tr>
<td>6.</td>
<td>Correct cluster numbers</td>
<td>20</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>火山</td>
<td>Total cluster numbers</td>
<td>28</td>
<td>33</td>
<td>Company name, Health, Area, Logevity..</td>
</tr>
<tr>
<td>7.</td>
<td>Correct cluster numbers</td>
<td>13</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>玄奘</td>
<td>Total cluster numbers</td>
<td>30</td>
<td>31</td>
<td>Nation name, Newspaper name, Shipbuilding..</td>
</tr>
<tr>
<td>8.</td>
<td>Correct cluster numbers</td>
<td>26</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>玉皇</td>
<td>Total cluster numbers</td>
<td>32</td>
<td>35</td>
<td>Cosmetic, Korean tradition dress, Cremation..</td>
</tr>
<tr>
<td>9.</td>
<td>Correct cluster numbers</td>
<td>27</td>
<td>29</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Number of initial clusters</th>
<th>2-depth</th>
<th>3-depth</th>
<th>4-depth</th>
<th>5-depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>68.73</td>
<td>72.54</td>
<td>73.12</td>
<td>72.14</td>
</tr>
<tr>
<td>10</td>
<td>75.42</td>
<td>86.74</td>
<td>87.04</td>
<td>82.48</td>
</tr>
<tr>
<td>15</td>
<td>70.90</td>
<td>77.29</td>
<td>75.24</td>
<td>76.68</td>
</tr>
<tr>
<td>20</td>
<td>65.32</td>
<td>71.38</td>
<td>73.38</td>
<td>70.49</td>
</tr>
<tr>
<td>25</td>
<td>63.49</td>
<td>70.12</td>
<td>64.67</td>
<td>62.38</td>
</tr>
<tr>
<td>30</td>
<td>54.21</td>
<td>68.89</td>
<td>60.43</td>
<td>58.49</td>
</tr>
</tbody>
</table>

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