Clustering Graduate Theses Based on Key Phrases Using Agglomerative Hierarchical Methods：An Experiment

Jau-Hwang Wang, Ju-Cheng Hsieh

Department of Information Management
Central Police University
Tao-Yuan, Taiwan
jwang@sun4.cpu.edu.tw

Abstract

Document clustering is an important tool for applications such as Web search engines. Document clustering can be defined as the process of organizing documents into groups. The groups thus formed have a high degree of association between members within the same group and a low degree of association between members of different groups. The goal of this paper is to present an experiment on one of the most widely used document clustering algorithms, namely, the *agglomerative hierarchical algorithm*. In our experiment, two set of graduate theses are clustered based on the key phrases assigned to each document by the author(s). Overall, the clustering results of our clustering scheme are considered to be very good.

Keywords: *Document Clustering, Agglomerative Hierarchical Clustering, Complete Link, Data Mining.*

1. Introduction

Document clustering is an important tool for applications such as Web search engines. The widely application of WEB technology has created a huge amount of web pages and the number of web pages is still increasing. According to an International Data Corporation report, the annual growth rate of storage media is more than 130%. Due to the huge amount of web pages, search engines have to be developed to help web users to search and retrieve information from the web in a timely fashion. As the number of web pages increases, the efficiency of web storage and retrieval becomes an important issue. Since the classification of web contents and the organization of web storage can have critical impacts on the retrieval performance of a search engine, some search engines, such as YAHOO, organize the web storage by means of laborious, time-consuming classification procedures. However, the accelerating influx of new web pages threatens to outpace the ability of human experts to classify the web contents. Therefore, automatic classification (also referred to as cluster analysis or clustering) methods must be developed to help alleviate this burden. Furthermore, search engines may return too many web pages for a particular key word search. Again clustering can be used to generate a category structure and enable users to have a better overview of the information contents [1].

Document clustering can be defined as the process of organizing documents into groups. The groups thus formed have a high degree of association between members within the same group and a low degree of association between members of different groups. While clustering is often referred to as automatic classification, it is not accurate strictly since the clusters formed are not known prior to processing, but are defined by the items assigned to them [2]. Clustering is useful to provide structure in large data sets, because it is not necessary for the clusters (and often the number of clusters) to be identified prior processing. Thus, it has been described as tool of discovery and also has been and important research area in data mining [3]. There are two major styles of clustering: partitioning (often called k-clustering), in which every document is assigned to exactly one group, and hierarchical clustering, in which each group of size greater than one is in turn consisted of smaller groups [2]. Both had been studied extensively by the mid-1970's, and comparatively less clustering research in the 1980's. However, the widely application of Web technology and the large amount of data thus created have lead to a renewal of interest in clustering algorithms. The goal of this paper is to present an experiment on one of the most widely used document clustering algorithms, namely, the agglomerative hierarchical algorithm.

Clustering can be performed on documents in several ways, such as clustering documents based on the terms that they contain, clustering documents based on co-occurring citations, and clustering terms based on the documents in which they co-occur [2]. In this experiment, two set of graduate theses from Taiwan are clustered based on the key phrases assigned to each document by their author(s).

The rest of the paper is organized as follows: section 2 describes the general agglomerative hierarchical clustering algorithm, section 3 describes the detail of our experiment, section 4 presents some results of the experiment, and the conclusions are given in section 5.

2. The Agglomerative Hierarchical Clustering Algorithm

The agglomerative hierarchical document clustering process includes the following three steps:

(1). Select the attributes for each document to be clustered. In principle, document clustering might involve a direct comparison of words or sentences

used in documents. However, the vocabularies of normal documents show substantial variety and the number of words or sentences included in many documents may be so large that a complete text comparison between different documents becomes impossible. Thus, it is advisable to characterize document by assigning special content descriptions, or profiles, which serve as document surrogates during cluster analysis [4]. The process of constructing identifiers as surrogates for documents is known as indexing. The choice of index terms should consider the degree that all aspects of the subject matter of a document are actually recognized and the index terms can somehow distinguish between different documents. Since indexing is rather a complex task, it was normally performed intellectually by subject experts, or by trained persons with experience in assigning content descriptions. It has been a routine for an author to assign key words for the document he/she has created. Thus, the key words assigned by the author(s) can be served as attributes for a document and used as the basis for cluster analysis.

(2). Select an appropriate similarity measure from those available. There are a variety of distance and similarity measures, such as *Simple Matching Coefficient, Dice Coefficient, Jaccard Coefficient, Overlap Coefficient,* and *Cosine Coefficient* [5,6]. A list of the similarity measures appears in Table 1.

Table 1. The Similarity Measures and Definitions

Similarity Measures	**Definitions, Sim(D_i, D_j) =**
Simple matching coefficient	$\lvert T_i \cap T_j \rvert$
Dice coefficient	$\frac{2\lvert T_i \cap T_j \rvert}{\lvert T_i \rvert + \lvert T_j \rvert}$
Jaccard coefficient	$\frac{\lvert T_i \cap T_j \rvert}{\lvert T_i \cup T_j \rvert}$
Overlap coefficient	$\frac{\lvert T_i \cap T_j \rvert}{\max(\lvert T_i \rvert, \lvert T_j \rvert)}$
Cosine	$\frac{\lvert T_i \cap T_j \rvert}{\sqrt{\lvert T_i \rvert \times \lvert T_j \rvert}}$

Where T_i is the set of terms assigned to document D_i, T_j is the set of terms assigned to document D_j, and $|\ T_i\ |$ is the number of elements of set T_i.

The Dice, Jaccard and cosine coefficients are three typical similarity measures, which have the attractions of simplicity and normalization and have often been used for document clustering [2]. The Jaccard is selected as the similarity measure for its simplicity in this experiment to calculate the similarity matrix for the initial data collection.

(3). Create the clusters or cluster hierarchies. Based on the similarity matrix, the two closet clusters are combined to form a new cluster. Once new clusters are created, the similarity matrix between clusters needed to be recalculated. The clustering process is repeated until a single cluster is obtained or there are no pairs of clusters having a similarity value larger than a predefined threshold. To calculate the similarity between clusters which have two or more members, four commonly used methods, namely, *single link, complete link, group average link,* and *Ward's method* can be used [2,7]. The clustering structure resulting from a hierarchical agglomerative clustering is often display as dendrogram as shown in Figure 1.

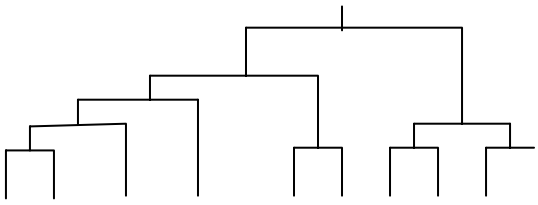


Figure 1. Dendrogram of a Hierarchical Clustering

These four methods are also called *maximum distance, minimum distance, group average distance,* and *centroid distance* respectively. Their definitions are as follows:

- *Minimum distance*：

$$d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'|$$

- *Maximum distance*：

$$d_{\max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} |p - p'|$$

- *Mean distance*： $d_{mean}(C_i, C_j) = |m_i - m_j|$
- *Average distance*：

$$d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'|$$

Where C_i or C_j represents a cluster, and p and p' are points (or members) of a cluster. Among the four typical measures, the complete link method has been shown to be most effective for larger collections [8] and is used for our experiment, since the size of document collection for the experiment is fairly large.

3. Experimental Details

The agglomerative hierarchical clustering algorithm used in this experiment can be summarized as the following steps:

(1). Initially assume that each document item forms a cluster.

(2). Calculate the similarity matrix for each pair of clusters using Jaccard Coefficient.

(3). Identify the two closest clusters and combine them in a cluster.

(4). Recalculate the similarity matrix for the newly created clusters using complete link method.

(5). If more than one cluster remains and there are some pairs of clusters whose similarity is greater than the threshold, which is set to 0 in our experiment, return to step (3).

The algorithm can be illustrated by example 1.

Example 1:

Step 1: Initially assume that each document item forms a cluster. Consider a document collection consists of ten documents, and the set of key phrases for the documents are $T_1, T_2, T_3, \ldots, T_{10}$ respectively. The key phrase sets are shown in Table 2.

Table 2. The Key Phrases Assigned to Documents

Documents	Key phrases
T_1	K1 K2 K3 K4
T_2	K5 K6 K7
T_3	K8 K9
T_4	K2 K3
T_5	K6 K7 K10
T_6	K8 K11 K12
T_7	K9 K11
T_8	K3 K4 K13 K14
T_9	K5 K6 K15 K16
T_{10}	K8 K11 K17 K18 K19

Let each document forms a cluster by itself, so there are ten clusters, denoted as $(T_1),(T_2),(T_3),\ldots,(T_{10})$.

Step 2: Calculate the similarity matrix for each pair of clusters. Using Jaccard coefficient, for example, the similarity between T_1 and T_4 is :

$$\text{Sim}(T_1,T_4) = \frac{|\ T_1 \cap T_4\ |}{|\ T_1 \cup T_4\ |} = \frac{|\ \{K2, K3\}\ |}{|\ \{K1, K2, K3, K4\}\ |} = \frac{2}{4} = 0.5.$$

Similarly, the similarity matrix, M_1, for the document collection can be calculated and shown in Figure 2.

$$M_1 = \begin{array}{c} \\ (T_1) \\ (T_2) \\ (T_3) \\ (T_4) \\ (T_5) \\ (T_6) \\ (T_7) \\ (T_8) \\ (T_9) \\ (T_{10}) \end{array} \begin{array}{c} \begin{array}{cccccccccc} (T_1) & (T_2) & (T_3) & (T_4) & (T_5) & (T_6) & (T_7) & (T_8) & (T_9) & (T_{10}) \end{array} \\ \begin{bmatrix} - & 0 & 0 & 0.5 & 0 & 0 & 0 & 0.33 & 0 & 0 \\ 0 & - & 0 & 0 & 0.5 & 0 & 0 & 0 & 0.4 & 0 \\ 0 & 0 & - & 0 & 0 & 0.25 & 0.33 & 0 & 0 & 0.17 \\ 0.5 & 0 & 0 & - & 0 & 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & - & 0 & 0 & 0 & 0.17 & 0 \\ 0 & 0 & 0.25 & 0 & 0 & - & 0.25 & 0 & 0 & 0.33 \\ 0 & 0 & 0.33 & 0 & 0 & 0.25 & - & 0 & 0 & 0.17 \\ 0.33 & 0 & 0 & 0.2 & 0 & 0 & 0 & - & 0 & 0 \\ 0 & 0.4 & 0 & 0 & 0.17 & 0 & 0 & 0 & - & 0 \\ 0 & 0 & 0.17 & 0 & 0 & 0.33 & 0.17 & 0 & 0 & - \end{bmatrix} \end{array}$$

Figure 2. The Similarity Matrix of M_1

Step 3: Identify the two closest clusters and combine them into a cluster. In this case, (T_1) and (T_4) are combined to form a new cluster (T_1, T_4), also (T_2) and (T_5) are combined to form another new cluster (T_2, T_5).

Step 4: Recalculate the similarity matrix for the newly created clusters using complete link method. The similarity matrix, M_2, for newly created clusters is shown in Figure 3.

$$M_2 = \begin{array}{c} \\ (T_1,T_4) \\ (T_2,T_5) \\ (T_3) \\ (T_6) \\ (T_7) \\ (T_8) \\ (T_9) \\ (T_{10}) \end{array} \begin{array}{c} \begin{array}{cccccccc} (T_1,T_4) & (T_2,T_5) & (T_3) & (T_6) & (T_7) & (T_8) & (T_9) & (T_{10}) \end{array} \\ \begin{bmatrix} - & 0 & 0 & 0 & 0 & 0.20 & 0 & 0 \\ 0 & - & 0 & 0 & 0 & 0 & 0.17 & 0 \\ 0 & 0 & - & 0.25 & 0.33 & 0 & 0 & 0.17 \\ 0 & 0 & 0.25 & - & 0.25 & 0 & 0 & 0.33 \\ 0 & 0 & 0.33 & 0.25 & - & 0 & 0 & 0.17 \\ 0.20 & 0 & 0 & 0 & 0 & - & 0 & 0 \\ 0 & 0.17 & 0 & 0 & 0 & 0 & - & 0 \\ 0 & 0 & 0.17 & 0.33 & 0.17 & 0 & 0 & - \end{bmatrix} \end{array}$$

Figure 3. The Similarity Matrix of M_2

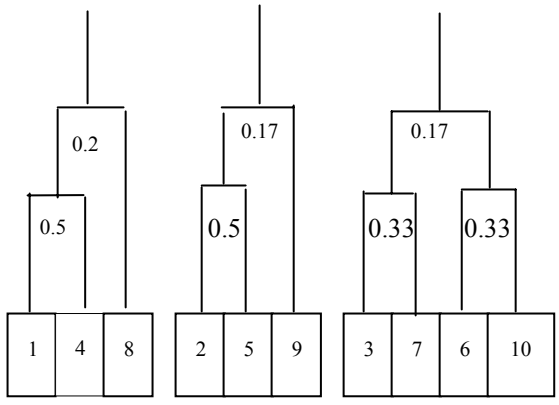
Step 5: If more than one cluster remains and there are some pairs of clusters whose similarity is greater than 0, return to step 3. In this case the algorithm returns to step 3.

The process repeats until the condition stated in step 5 is not true and eventually the final similarity matrix, M_f, is created and shown in Figure 4.

$$\begin{array}{c} \\ (T_1,T_4,T_8) \\ (T_2,T_5,T_9) \\ (T_3,T_6,T_7,T_{10}) \end{array} \begin{array}{c} \begin{array}{ccc} (T_1,T_4,T_8) & (T_2,T_5,T_9) & (T_3,T_6,T_7,T_{10}) \end{array} \\ \begin{bmatrix} - & 0 & 0 \\ 0 & - & 0 \\ 0 & 0 & - \end{bmatrix} \end{array}$$

Figure 4. The Final Similarity Matrix of M_f

Finally, the clusters (T_1, T_4, T_8)、(T_2, T_5, T_9)、(T_3, T_6, T_7, T_{10}) are created and the clustering process is shown in Figure 5.



(T_1) (T_4) (T_8) (T_2) (T_5) (T_9) (T_3) (T_7) (T_6) (T_{10})

Figure 5. The Agglomerative Hierarchical Document Clustering Process

4. Experimental Results

Two data sets of graduate theses published by universities in Taiwan are used as clustering targets in this experiment. The first data set consists of 411 master theses published by 8 departments[1] from Central Police University. The second data set is the graduate theses published in Taiwan and is a much bigger collection, but due to computation time limit only 1078 master theses published by 5 different departments[2] in Taiwan are used for the experiment. The clustering of the first data set shows that most theses published by an academic department form one single cluster. Only the theses published by the Department of Police Administration are clustered into two different clusters. So totally, 9 clusters are created. Although, the theses published by Police Administration Department form two clusters, 98.56% of them are clustered into one cluster.

The number of clusters created from clustering the second data set is 36, which is much bigger than 5. However, 89.4% of atmospheric science theses, 90.3% of marine biology theses, 89.6% of international economics theses, 90% of plant pathology theses, and 86.7% of electro-physics theses, are clusters into five main clusters respectively. Overall, the clustering results are considered to be very good.

5. Conclusions

In this paper, we have described the agglomerative hierarchical clustering method in detail and the experiments to cluster two collections of graduate theses. It is shown that based on the key phrases assigned to documents by the author(s), the agglomerative hierarchical clustering method is able to cluster most (about 90%) of the theses published in one academic area to a single clusters. For theses published by Central Police University, the clustering result is much better, and this may due to that the research area of Police University is more specific and better focused.

Since the academic departments are often used as the categories for classification, we conclude that our clustering scheme is promising for automatic document classification if the clustering granularity is on the academic department basis. For clustering documents on other levels of granularity, terms or individual words besides key phrases and also associated their weights might be used as surrogates for documents. In that case the clustering scheme will be more complex to implement and thus more computation efforts is needed.

---

[1] They are police administration, fire science, criminal police, traffic administration, information management, crime prevention, forensic science, and law.

[2] They are atmospheric science, marine biology, international economics, plant pathology, and electro-physics.

References

[1] K. I. Lin, and R. Kondadadi, “A Word-Based Soft Clustering Algorithm for Documents”, www.msci.memphis.edu/~linki/_mypaper/CATA01.doc

[2] W. B. Frakes, and R. Baeza-Yates, *Information Retrieval: Data Structure & Algorithms*, Prentice Hall, 1992.

[3] D. Fasulo, “An analysis of recent work on clustering algorithms”, http://www.cs.washington.edu/homes/dfasulo /clustering.ps , 1999.

[4] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer,* pp-275, Addison Wesley, 1989.

[5] M. R. Anderberg, *Cluster Analysis for Applications,* Academic, 1973.

[6] 彭怡菁，*以統計量測爲基礎之交易資料集分群*，Master Thesis, National Taiwan University, 2001.

[7] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, August 2000.

[8] E. M. Voorhees, “The effectiveness and Efficiency of Agglomerative Hierarchic Clustering in Document Retrieval”, *Ph. D. Thesis*, Cornell University, 1986.

[9] M. Berry and G. Linoff, *Data Mining Techniques for Marketing, Sales, and Customer Support*, John Wiley and Sons, 1997.

[10] A.K. Jain, M.N. Murty and P.J. Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, Vol. 31, No. 3, pp.264-323, 1999.

[11] G. Jones, A. M. Robertson, C. Santimetvirul, and P. Willett, “Non-Hierarchic Document Clustering Using A Genetic Algorithm”, http://informationr.net/ir/1-1/paper1.html.