Chapter 21
Concepts Labeling of Document Clusters Using a Hierarchical Agglomerative Clustering (HAC) Technique

Rayner Alfred, Tan Soo Fun, Asni Tahir, Chin Kim On and Patricia Anthony

Abstract The most common way to organize and label documents is to group similar documents into clusters. Normally, the assumed number of clusters may be unreliable since the nature of the grouping structures among the data is unknown before processing and thus the partitioning methods would not predict the structures of the data very well. Hierarchical clustering has been chosen to solve this problem by which they provide data-views at different levels of abstraction, making them ideal for people to visualize the concepts generated and interactively explore large document collections. The appropriate method of combining two different clusters to form a single cluster needs affects the quality of clusters produced. In order to perform this task, various distance methods will be studied in order to cluster documents by using the hierarchical agglomerative clustering. Clusters very often include sub-clusters, and the hierarchical structure is indeed a natural constraint on the underlying application domain. In order to manage and organize documents effectively, similar documents will be merged to form clusters. Each document is represented by one or more concepts. In this paper, concepts that characterize English documents will be generated by using the hierarchical agglomerative clustering. One of the advantages of using hierarchical

R. Alfred (✉) · T. S. Fun · A. Tahir · C. K. On
School of Engineering and Information Technology, Center of Excellence in Semantic Agents, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
e-mail: ralfred@ums.edu.my

T. S. Fun
e-mail: soofun@ums.edu.my

A. Tahir
e-mail: asnieta@ums.edu.my

C. K. On
e-mail: kimonchin@ums.edu.my

P. Anthony
Faculty of Environment, Society and Design, Department of Applied Computing, Lincoln University, Christchurch, New Zealand
e-mail: patricia.anthony@lincoln.ac.nz

L. Uden et al. (eds.), *The 8th International Conference on Knowledge Management in Organizations*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7287-8_21,

clustering is that the overlapping clusters can be formed and concepts can be generated based on the contents of each cluster. The quality of clusters produced is also investigated by using different distance measures.

Keywords Hierarchical agglomerative clustering · Concepts aggregation · Automatic document labeling · Distance measure · Knowledge management

21.1 Introduction

The dramatic rise in the use of the web and the improvement in communications in general have transformed our society into one that strongly depends on information [4, 5]. Vast amounts of text documents are also available in various fields. The huge amount of various documents accumulates daily in databases on web are astonishing. The accumulations of available text documents have raised new challenges for information retrieval (IR) technology. Therefore, in order to facilitate the knowledge management process, various approaches and techniques applied on text classification (categorization) and text clustering are being compared and studied. In short, it is essential and important for us to manage the unstructured and random documents by labeling these documents automatically.

This paper proposes a novel approach to manage English text documents by clustering and labeling them automatically. Clustering is a frequent performed task and technique for machine learning, data mining, pattern recognition, image analysis and bioinformatics. It can be applied in various type of tasks related to improving the quality in the structure and usage of large and high dimensional data. It is a method of unsupervised learning in which a descriptive task will be performed. There are many potential applications and advantages that will accrue from being able to reliably and automatically cluster, categorize and label corpora of documents. Many of these document clustering works are based on supervised or unsupervised learning techniques in order to label particular web documents as belonging to a specific category, or grouping together similar documents into clusters.

In this paper, an unsupervised learning technique will be used to implement the proposed algorithm in which a hierarchical clustering is applied to unlabeled documents. The hierarchical clustering is chosen instead of partitional clustering (K-means) [3] because the hierarchical clustering is able to form overlapping clusters which is more suitable for this research. The taxonomy (tree) is able to show the sub-clusters of the parent cluster and thus the aggregation of concepts can be illustrated [9]. The two types of hierarchical clustering are agglomerative (bottom-up) and divisive. Hierarchical agglomerative clustering is applied in this research since the concepts obtained from sub-clusters can be aggregated [1]. Distance measures such as single linkage and complete linkage will also be compared in order to compare the results of clusters based on different distance methods.

This is paper is organized as followed. Section 2 introduces some related works and we present the details of our approach and our dataset. Section 3 describes the concept labeling of document clusters by using the hierarchical agglomerative clustering technique. Section 4 presents the experimental design set-up and the experimental results. This paper presents the evaluation of several structures of hierarchical agglomeration clustering results that are produced by using different types of inter-clusters distance measurements, namely Single, Complete and Average links. This paper is concluded with future works in Sect. 5.

21.2 Related Works

Earlier works include the comparison of Chi Square (X2), most frequent words, most predictive words and a combination of most frequent and predictive words methods [6]. In this comparison, Popescul and Ungar found that the most frequent and predictive words method produced the best labels, capturing the words which both occur frequently in a cluster and effectively discriminate the given cluster from the other clusters. Unfortunately, none of the methods gave uniformly satisfactory results when disciplines corresponding to clusters are very diverse in the vocabulary used and encompass very broad topics.

In other works, Lamirel et al. present a new approach combining original hypertree construction techniques for multidimensional clustering results visualization with novel cluster labeling techniques based on the use of cluster content evaluation criteria, like the F-measure on cluster properties [7]. Treeratpituk and Callan proposed a simple linear model that considers the structure of the hierarchy when automatically assigning labels to document clusters in a hierarchy [8]. They conducted a study to show the effectiveness of different statistical features in selecting cluster labels. They also showed that such a simple model is likely to tolerate the type of noise in the cluster hierarchy that is normally generated by clustering algorithms.

Newman et al. [9] explored visualizations of document collections, which they call topic maps. In their works, they found that while topics are a useful way to organize an entire collection, producing a static global topic map of the collection may have limited value for exploring the collection. Therefore local topic maps may ultimately be more useful for better understanding and navigating local structure in a collection.

Magatti et al. addressed the task of labeling topics that are induced by a hierarchical topic model [10]. Their label candidate set is the Google Directory (gDir) hierarchy, and label selection takes the form of ontological alignment with gDir. The disadvantages of this method are that the method is only applicable to a hierarchical topic model and crucially relies on a pre-existing ontology and the class labels contained therein.

Automatic labeling of document clusters can also be associated with the problem of topic extraction. The problem of topic extraction is attracting a great

deal of attention due to its wide applicability; extraction of scientific research topics, author-topic analysis, opinion extraction and information retrieval. Several probabilistic models have proved to be effective to discover topics [1–3].

In this paper, the authors propose a framework for concept aggregation based on hierarchical agglomerative clustering. This paper explores and describes the interplay between inter-cluster distance methods and the aggregated concepts extracted from the hierarchical concept model.

21.3 Concepts Labeling of Document Clusters

The document labeling problem can be tackled in two different ways: human labeling and computer labeling. The first approach maps a document into a set of pre-specified categories, usually such categories form a taxonomy or a topic hierarchy. The latter approach has recently emerged to be well suited, i.e. to be efficient and effective, in several settings. While human labeling usually benefits from the availability of a domain specific topic hierarchy, agreed by experts, it is extremely time consuming and in some particular situations universally agreed labeling cannot be achieved. On the contrary, computer labeling is economically attractive, while the achieved labeling must be accurately checked by specific domain experts to ensure that it is consistent. Furthermore, effective methods for automatically building a hierarchy of topics have been very recently proposed in the specialized literature [17]. The approach we propose offers an efficient way to extract concepts automatically from the document clusters. We apply the hierarchical agglomerative clustering technique to group documents. Hierarchical clustering has been chosen to solve this problem by which they provide data-views at different levels of abstraction, making them ideal for people to visualize the concepts generated and interactively explore large document collections.

21.3.1 Clustering Parallel Corpora

In this experiment, we use the vector space model [12], in which a document is represented as a vector in n-dimensional space (where n is the number of different words in the collection). Here, documents are categorized by the words they contain and their frequency. Before obtaining the weights for all the terms extracted from these documents, stemming and stopword removal is performed. Stopword removal eliminates unwanted terms (e.g., those from the closed vocabulary) and thus reduces the number of dimensions in the term-space. Once these two steps are completed, the frequency of each term across the corpus is counted and weighted using term frequency—inverse document frequency (*tf-idf*) [12], as described in Eq. (21.1).

$$tf - idf = tf(t,d) \cdot idf(t) \tag{21.1}$$

$$idf(t) = \log\left(\frac{|\mathrm{D}|}{\mathrm{df(t)}}\right) \tag{21.2}$$

$$\mathrm{sim}\left(d_i, d_j\right) = \frac{\left(\mathrm{d}_i \mathrm{d}_j\right)}{\left(\|\mathrm{d}_i\| \cdot \|\mathrm{d}j\|\right)} \tag{21.3}$$

Weights are assigned to give an indication of the importance of a word in characterizing a document as distinct from the rest of the corpus. In summary, each document is viewed as a vector whose dimensions correspond to words or terms extracted from the document. The component magnitudes of the vector are the *tf−idf* weights of the terms. In this model, *tf−idf*, as described in Eq. (21.1), is the product of term frequency *tf*(*t*,*d*), which is the number of times term *t* occurs in document *d*, and the inverse document frequency, Eq. (21.2), where |D| is the number of documents in the complete collection and *df*(*t*) is the number of documents in which term *t* occurs at least once. To account for documents of different lengths, the length of each document vector is normalized so that it is of unit length [13].

21.3.2 Hierarchical Agglomerative Clustering

In this work, we concentrate on hierarchical agglomerative clustering. Unlike partitional clustering algorithms that build a hierarchical solution from top to bottom, repeatedly splitting existing clusters, agglomerative algorithms build the solution by initially assigning each document to its own cluster and then repeatedly selecting and merging pairs of clusters, to obtain a single all-inclusive cluster, generating the cluster tree from leaves to root [11]. The main parameters in agglomerative algorithms are the metric used to compute the similarity of documents and the method used to determine the pair of clusters to be merged at each step.

In these experiments, the cosine distance, Eq. (21.3), is used to compute the similarity between two documents $\mathrm{d_i}$ and $\mathrm{d_j}$. This widely utilized document similarity measure becomes one if the documents are identical, and zero if they share no words. The two clusters to merge at each step are found using either, the single link, complete link or average link method. In this scheme, the two clusters to

Table 21.1 Five categories of english news extracted from the star online news

Categories	Number of text documents	Average words
Business	1550	315
Entertainment	610	753
Generals	1560	453
Politics	380	785
Sports	340	459

merge are those with the greatest minimum (Single link), maximum (Complete link) or average (Average link) similarity distances between the documents in one cluster and those in the other [15, 16].

21.3.3 Extracting Concepts of Document Clusters

Concepts that characterize English documents will be generated by using the hierarchical agglomerative clustering. This is done by computing terms that have large weights assigned to them to indicate the importance of a word in characterizing a document as distinct from the rest of the corpus.

21.4 Experimental Design and Evaluations

The experiment is designed in order to investigate and compare the effectiveness of clustering English text documents using three different types of inter-cluster distance measurement, namely minimum (*Single Link*), maximum (*Complete Link*) and average (*Average Link*). Depending on the type of inter-cluster distance used, cluster result that provides the lowest DBI value will be taken into consideration for the extraction of concepts to characterize the document clusters. There are five categories of English news collected from the Star Online in the year of 2010 (Malaysian local online news—thestar.com.my). The details of the text documents used in this experiment are shown in Table 21.1.

There are two main stages in this experiment that includes (a) Clustering English Documents using a Hierarchical Agglomerative Clustering technique and (b) Concepts Extraction.

21.4.1 Clustering English Documents Using a Hierarchical Agglomerative Clustering Technique

In the first stage, we perform the task of clustering English texts by using the Hierarchical Agglomerative Clustering. We look at the similarities of pair of clusters based on three inter-distance methods (*Single*, *Complete* and *Average* links). We evaluate the structure of tree-like cluster results that minimizes some objective function applied to *k*-cluster centers. In our case, we consider the *cluster dispersion.*

$$d_{centroid}(Q_k) = \frac{\sum_i \|xi - c_k\|}{N_k} \tag{21.4}$$

$$c_k = 1/N_k \sum_{xi \in Q_k} xi \tag{21.5}$$

$$d_{between}(Q_k, Q_l) = \|c_k - cl\| \tag{21.6}$$

$$DBI = \frac{1}{K}\sum_{k=1}^{K} \max l \neq k \left\{ \frac{dcentroid(Q_k) + dcentroid(Q_l)}{dbetween(Q_k, Q_l)} \right\} \tag{21.7}$$

Typically *cluster dispersion metric* is used, such as the Davies-Bouldin Index (DBI) [14]. DBI uses both the intra-cluster and inter-clusters distances to measure the cluster quality. Let $d_{centroid}(Q_k)$, defined in (4), denotes the average link distances within-cluster Q_k, where $x_i \in Q_k$, N_k is the number of samples in cluster Q_k, c_k is the center of the cluster and $k \leq K$ clusters. Let $d_{between}(Q_k, Q_l)$, defined in (6), denotes the distances inter-clusters Q_k and Q_l, where c_k is the centroid of cluster Q_k and c_l is the centroid of cluster Q_l. In this study, we also cluster the text documents based on the minimum (single link), maximum (complete link) and average (Average link) distances between clusters. Therefore, given a partition of the N points into K-clusters, DBI is defined in (7). This *cluster dispersion* measure can be incorporated into any clustering algorithm to evaluate a particular segmentation of data.

Table 21.2 Comparison of DBI values for the clustering results when using single, complete and average links with different number of clusters, k = 5, 10 and 15

Categories	k	DBI		
		Single link	Complete link	Average link
Business	5	29.1	28.1	28.4
	10	29.9	28.2	28.3
	15	30.4	28.7	28.1
	Average	**29.8**	**28.3**	**28.3**
Entertainment	5	35.2	31.1	32.8
	10	36.0	30.2	33.8
	15	36.5	32.3	33.9
	Average	**35.9**	**31.2**	**33.5**
Generals	5	29.1	27.5	28.2
	10	28.9	27.2	28.8
	15	30.5	28.6	29.6
	Average	**29.5**	**27.8**	**28.9**
Politics	5	34.2	30.1	33.8
	10	35.0	30.2	34.8
	15	35.7	31.3	34.9
	Average	**34.9**	**30.5**	**34.5**
Sports	5	28.2	28.1	28.1
	10	28.9	28.1	28.9
	15	29.2	28.6	29.1
	Average	**28.7**	**28.3**	**28.7**

Table 21.3 Comparison of actual concepts and extracted concepts for all five news categories for k = 10

Categories	Actual concepts	Extracted concepts	% of concepts aligned
Business	**car**, **oil industry**, **income tax**, market,	**income**, high, country, **proton**, **petrona**	60
Entertainment	industry, idol, **artist**, life style, **tourism**	film, **tourist**, **artist**, vacat, trip	40
Generals	**transportation**, **cabinet appointment**, economy, weather	**transport**, ministry, long, secretary, **appoint**	40
Politics	**election**, government, opposition, **minister**, **voting**	najib, indian, **voter, regist, minist**	60
Sports	**badminton**, **open**, championship, **soccer**, bowling	play, set, **chong**, round, **open**, **football**	60

21.4.2 Concepts Extraction

Next, in the second stage, we compute the weights of terms (*tf*−*idf*) considered in clustering English documents for each cluster. The top five terms with high weights for each cluster will be extracted as concepts that characterize each corresponding cluster. The extracted concepts are then compared with the actual concepts derived manually.

21.4.3 Experimental Results

Table 21.2 shows the comparison of DBI values for the clustering results using *Single*, *Complete* and *Average* links with different number of clusters produced. Based on the results obtained, on average the quality of clusters produced is better for documents when the *Complete* link is used as the inter-cluster distance in the process of clustering them. This is due to the fact that when the *Complete* link method is used to cluster English documents, two clusters, C_i and C_j having two elements, e_a and e_b where $e_a \in C_i$ and $e_b \in C_j$, with the highest value of distance are merged into one cluster. As a result, the final clusters produced may be well separated and thus produces lower DBI value. The results also show that the quality of clusters produced is better when we have smaller number of clusters. Another important finding that can be obtained from this experiment is that when the number of words is high, the DBI values are also high as shown in Table 21.2 for the Entertainment and Politics categories.

Table 21.3 shows the concepts extracted from English clusters using the hierarchical clustering technique with $k = 10$ and Complete link method. We compare

the actual concepts derived manually and the extracted concepts by referring to the tree-like document representation and computing the top five terms based on the *tf*−*idf* weights. Based on the results shown in Table 21.3, the percentage of concepts aligned between the actual concepts derived manually and the extracted concepts derived from the HAC technique is quite encouraging, with the exception of the Entertainment and Generals news categories. This is probably because the Entertainment and Generals news are very diverse in the vocabulary used and encompass very broad topics.

21.5 Conclusion

In this paper we have presented the framework of using a tree-like document clusters representation which is produced by clustering English documents using the Hierarchical Agglomerative Clustering (HAC) technique to automatically extract concepts that characterize each cluster. We have empirically showed that better clustering results can be produced by using the *Complete* link method in computing the inter-cluster distance measurement when merging two different clusters. By using the *Complete* link method to merge two different clusters, the concepts extracted from each cluster should be more relevant and applicable in labeling the English clusters as shown in the experimental results. In order to improve the results obtained, future works include comparing actual concepts with extracted concepts derived based on other weights computation and implementing a semi-supervised HAC technique to extract concepts automatically from English document clusters more effectively.

Acknowledgments This work has been supported by the Long Term Research Grant Scheme (LRGS) project funded by the Ministry of Higher Education (MoHE), Malaysia under Grants No. LRGS/TD/2011/UiTM/ICT/04.

References

1. Nasser AMB, Jian-Hui J, Ru-Qin Y (2005) Bubble agglomeration algorithm for unsupervised classification: a new clustering methodology without a priori information. Chemometr Intell Lab Syst 77(1–2):43–49
2. Reynaldo GG, Aurora PP (2010) Dynamic hierarchical algorithms for document clustering. Pattern Recogn Lett 31(6):469–477
3. Xiaojun W, Jianwu Y (2007) CollabSum: exploiting multiple document clustering for collaborative single document summarizations. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrievalm, Amsterdam, The Netherlands, 23–27 July 2007, pp 143–150
4. Carullo M, Binaghi E, Gallo I (2009) An online document clustering technique for short web contents. Pattern Recogn Lett 30:870–876

5. Iliopoulos I, Enright AJ, Ouzounis CA (2001) Textquest: document clustering of medline abstracts for concept discovery in molecular biology. Pac Symp Biocomput 6:384–395
6. Popescul A, Ungar LH (2000) Automatic labeling of document clusters. http://citeseer.nj.nec.com/popescul00automatic.html
7. Lamirel JV, Ta AP, Attik M (2008) Novel labeling strategies for hierarchical representation of multidimensional data analysis results. In: Gammerman A (ed) Proceedings of the 26th IASTED international conference on artificial intelligence and applications (AIA '08). ACTA Press, Anaheim, pp 169–174
8. Treeratpituk P, Callan J (2006) An experimental study on automatically labeling hierarchical clusters using statistical features. SIGIR 2006:707–708
9. Newman, Baldwin T, Cavedon L, Karimi S, Martinez D, Zobel J (2010) Visualizing document collections and search results using topic mapping. J Web Semant 8(2–3):169–175
10. Magatti, Calegari S, Ciucci D, Stella F (2009) Automatic labeling of topics. In: ISDA 2009, Pisa, pp 1227–1232
11. Zhao Y, Karypis G (2005) Hierarchical clustering algorithms for document datasets. Data Mining Knowl Disc 10(2):141–168
12. Salton G, Michael JM (1986) Introduction to modern information retrieval. McGraw-Hill Inc., New York
13. van Rijsbergen CJ (1979) Information retrieval, 2nd edn. Butterworths, London
14. Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1(2):224–227
15. Khalilian M, Mustapha N Data (2010) Stream clustering: challenges and issues, In: Proceedings of the international multiconference of engineers and computer scientists IMECS, Hong Kong, pp 978–988
16. Torres GJ, Basnet RB, Sung AH, Mukkamala S, Ribeiro BM (2009) A similarity measure for clustering and its applications. Int J Electr Comput Syst Eng 3(3):164–170
17. Alfred R, Kazakov D, Bartlett M, Paskaleva E (2007) Hierarchical agglomerative clustering for cross-language information retrieval. Int J Transl 19(1):139–162