A Phrase-Based Method for Hierarchical Clustering of Web Snippets

Zhao Li and **Xindong Wu**
Department of Computer Science
University of Vermont
{zhaoli,xwu}@cems.uvm.edu

Abstract

Document clustering has been applied in web information retrieval, which facilitates users' quick browsing by organizing retrieved results into different groups. Meanwhile, a tree-like hierarchical structure is well-suited for organizing the retrieved results in favor of web users. In this regard, we introduce a new method for hierarchical clustering of web snippets by exploiting a phrase-based document index. In our method, a hierarchy of web snippets is built based on phrases instead of all snippets, and the snippets are then assigned to the corresponding clusters consisting of phrases. We show that, as opposed to the traditional hierarchical clustering, our method not only presents meaningful cluster labels but also improves clustering performance.

Introduction

Requested by a query, most of search engines return the retrieved results in a ranked list. It is a time-consuming task for users to find what they are looking for, especially when they are unfamiliar with the topics relevant to the query. Clustering web snippets (a list of retrieved results consisting of titles and abstracts) into distinct topics enables the uses to quickly identify their required information. Some traditional clustering techniques have been investigated in this regard, following the way of grouping web snippets into clusters based on similarity measurements and then generating labels for clusters by extracting words of high occurring frequency in the snippets. However, these individual words are not able to represent the clusters, which makes it still difficult for the users to effectively browse the topics of interest. To solve this problem, a Suffix Tree Clustering (STC) was first proposed in (Zamir and Etzioni 1998). Based on the suffix tree data structure, STC first extracts common phrases shared by documents as candidate cluster labels. The documents are assigned to the relevant phrases to form candidate clusters, and the final clusters are obtained by merging overlapping clusters which share a certain number of documents. In (Zeng et al. 2004), these phrases were presented as a ranking list, based on a regression model learned from human labeled training data. These methods are effective for clustering web snippets in the sense of describing the clusters; because they are based on the fact that phrases, rather than single words, are more useful in constructing an intuitive description of clusters. However, the methods discussed above share a common disadvantage in determining the number of clusters: selecting top k clusters or asking uses to specify. Finding the right number of clusters in a set of documents is very difficult when one does not exactly know the information needs of the users. This problem would be circumvented by using a hierarchy of clusters rather than clusters residing at the same level. A hierarchical agglomerative clustering algorithm (Zhao and Karypis 2002) starts with all documents belonging to their individual clusters and combines the most similar clusters until the desired number of clusters is obtained. By describing the relationship between groups of documents one would browse at a specified level. However, a basic hierarchical agglomerative clustering algorithm is also subject to generating meaningful cluster labels. In this paper, we introduce a new method for clustering web snippets. First, it extracts all salient phrases and builds a phrase-based document index; Then it starts with all extracted phrases belonging to their individual clusters and combines the most similar clusters by using the phrase-based document index; Finally, each cluster is identified by a distinct phrase, and a snippet is assigned to a cluster, in which all of the indexing phrases of the snippet belong to. For those snippets whose consisting phrases belong to different clusters are identified by their assigned neighbors. Our method can generate meaningful cluster labels, and the experiments show that our method can effectively improve clustering performance.

Phrase-based Hierarchical Clustering

Most of hierarchical clustering algorithms compute the pairwise similarity between documents. Mostly, a hierarchical tree is built with several levels of no description, which makes it difficult to find the concept-wise relationship of groups. In nature, phrases are better suited to represent the relationship of groups. It is worthy of noting that documents can be treated as a collection of phrases. Documents belonging to the same cluster share a certain number of common phrases, which can be used to identify each cluster. In the case of clustering web snippets, the number of phrases is far less than that of snippets, which can be used to construct a concise hierarchy tree of concepts by using a phrase-based

Table 1: *Summary of datasets*

query	# classes	# docs	# words	# phrases
topology	4	89	169	18
geometry	5	197	364	54
number theory	7	288	470	70
meat	8	352	442	150
knowledge	9	831	1186	266

document index. Our phrase-based hierarchical clustering algorithm is described as follows:

```
Extract all the $n$-grams as candidate phrases $P = \{p_1, p_2, ..., p_l\}$, based on suffix tree built from a collection of web snippets $D = \{d_1, d_2, ..., d_m\}$.
Build a phrase-based document index $R = \{r_{p_1}, r_{p_2}, ..., r_{p_l}\}$, where $r_{p_k}$ contains the indexed documents of $p_k$.
Construct a Vector Space Model of $D$ based on $tf-idf$ measurement, and calculate the proximity matrix for phrases by using group average technique: $proximity(p_i, p_j) = \frac{\sum_{d_i \in r_{p_i}, d_j \in r_{p_j}} proximity(d_i, d_j)}{|r_{p_i}| * |r_{p_j}|}$.
while the desired number of clusters is not obtained do
    Merge the closest two clusters, and select a phrase of highest number of indexing documents from the merging clusters as the new cluster label.
    Update the proximity matrix between the new cluster and the original clusters.
end while
Assign the snippets whose indexing phrases belong to the same cluster.
Assign the remaining snippets based on their $k$-nearest assigned neighbors.
```

Experiments

We selected five hierarchically well-organized datasets from the Open Document Project (http://www.dmoz.org/), which is a large human-edited hierarchical directory of the Web. The number of n-grams of phrase was set as $2 \leq n \leq 5$. Stop words and unique indexing phrases were filtered. Moreover, the phrases of frequency less than 3, and the words of frequency less than 2 were ignored. A summary of these dataset is listed in Table 1. For these snippets assigned by their k-nearest neighbors, k was denoted by 5. We evaluate our method (PHAC) with a baseline hierarchical agglomerative clustering algorithm (HAC) in terms of two popular metrics: F-measure and entropy. F-measure is an aggregation of precision and recall. The higher of F-measure, the better of clustering. Entropy tells how homogeneous a cluster is. The higher the homogeneity of a cluster, the lower the entropy is, and vice versa. The results are shown in Figure. 1 and Figure. 2.

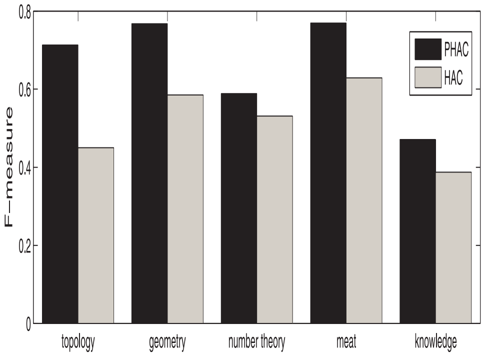

Figure 1: *Clustering quality comparison of F-measure*

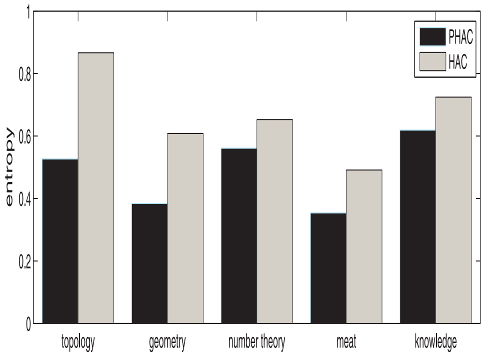

Figure 2: *Clustering quality comparison of entropy*

Conclusions

We have introduced a new method for hierarchical clustering of web snippets in this paper, which is based on the fact that similar web snippets share a small amount of phrases. By constructing a hierarchy of phrases, each cluster is labeled by a distinct phrase, and the snippets can be clustered by utilizing the phrase-based document index. Experiments showed that our method outperforms the traditional hierarchical clustering algorithm.

References

Zamir, O., and Etzioni, O. 1998. Web document clustering: a feasibility demonstration. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*, 46–54. New York, NY, USA: ACM.

Zeng, H.-J.; He, Q.-C.; Chen, Z.; Ma, W.-Y.; and Ma, J. 2004. Learning to cluster web search results. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval*, 210–217. New York, NY, USA: ACM.

Zhao, Y., and Karypis, G. 2002. Evaluation of hierarchical clustering algorithms for document datasets. In *CIKM '02: Proceedings of the eleventh international conference on information and knowledge management*, 515–524.