D-038

Hierarchical Clustering and Bisecting K-Means in producing Time Series Patent Map

Maryjane ANDAL Shigeru OYANAGI Katsuhiro YAMAZAKI Masatoshi KAMIHARAKO

Graduate School of Science and Engineering, Ritsumeikan University
Biwako-Kusatsu Campus Noji Higashi
1 chome, 1-1 Kusatsu, 525-8577 Shiga-ken, JAPAN

I Introduction

Clustering documents into meaningful groups helps on providing fast and easy way of exploring documents collections. This may could even lead to discovery of new knowledge.

Consequently, researches on clustering techniques are still emerging nowadays, either for evaluation, improvement or potential application. And this paper revisits Hierarchical Clustering techniques and K-Means, the two well-known techniques in clustering and looks for its potential application in patent documents. However, the research attempts only to produce time series map of a particular technology field, which is one type of many patent maps that can be generated from patent document collections.

This paper discusses Hierarchical Clustering Techniques particularly Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and Bisecting K-Means. Then, propose a framework that combines these two methods in producing a time series patent map. Datasets used in figures and illustrations in this paper are patent document collections downloaded from USPTO Website.

II UPGMA and Bisecting K-Means

Hierarchical Clustering is one method in cluster analysis or data segmentation used in text mining. It clusters documents based on its similarity and organizes in a form of hierarchy [3]. Similarity is decided by the average (UPGMA), minimum (Single Linkage) or maximum (Complete Linkage) distances between documents within a cluster [6]. These are known as agglomerative techniques. Among these three, UPGMA outperformed the other two techniques [1,2] using variety of small to bigger datasets. In figure 1 also, which uses patent abstracts shows that UPGMA on the average result is more stable and outperforms single linkage and complete linkage, even for a very few documents.

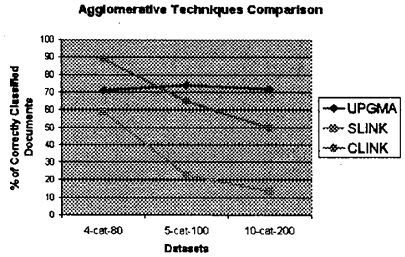

Figure 1 *Agglomerative Techniques* Comparison

However, still, it does not yield a very high percentage of correctly classified documents. This may be due to its known problem in agglomeration, which is chaining phenomenon wherein if the documents are not part of a particular cohesive groups, and initial merging decision may contain some errors, this tends to be multiplied as agglomeration progresses [2]. To reduce error in agglomeration at early stage, it has been proposed to performed partitional clustering algorithm (via repeated bisection of K-means approach) first to discover clusters then perform agglomeration to form a hierarchy. This is termed as *constrained agglomerative algorithm* [2]. It was found out that constraining leads to purer neighborhoods as it can identify the right set of dimensions for the various clusters [2].

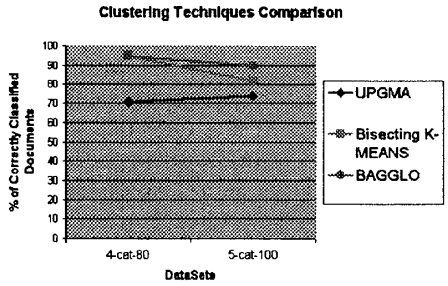

Figure 2 *Clustering Techniques Comparison*

Zhao and Karpyis [2] presents this algorithm using eleven datasets from various sources, it shows that it outperformed agglomerative method and for many cases also with the partitional algorithm. This is same with the result using patent abstracts, which is shown in Figure 2. They also referred to this method as biased agglomerative (BAgglo) clustering algorithm, which uses partitional clustering solution via repeated bisecting K-means approach to bias the agglomeration process. Hence, this paper uses this approach using CLUTO (Clustering Toolkit) [4] to cluster patent documents and map in time series.

III Application to Time Series Patent Map

Clustering patent documents under same international patent classification is exploring on a particular technology field, that could either give an idea on it's technology progress across time or relationship between each clustered patent documents.

This had been illustrated using the proposed framework shown in Figure 3. UPGMA and Bisecting K-Means were used in clustering searched patent documents and finding relationships among clustered documents.

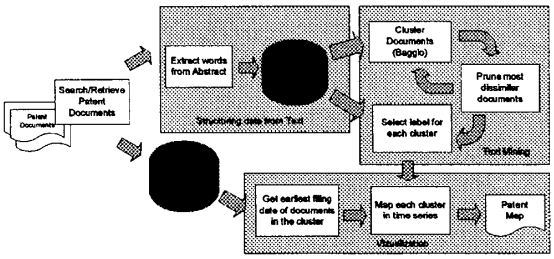

Figure 3 *Proposed Framework in Producing Time Series Map*

At first, think on what particular technology field one would like to explore. In this paper, 39 retrieved patent documents related to payment technology are used. Then, words are extracted from patent abstracts and brief technology description of each document. Brief technology description had been included in word extraction to better provide words that would best describe patent document. Since some patent abstracts are too small for word extraction. Using TF/IDF formula, corresponding weight for each word had been computed.

Then, perform clustering. From the tree produced, a patent expert may want to exclude some clusters, which are not that similar to the topics but had been included because of simple search that was done. Percentage of similarities among clusters and among documents within the clusters could help the patent expert on deciding which clusters could be pruned from the tree. After, selecting the final clusters to retain for visualization, the patent expert may label each cluster through selecting among descriptive words displayed in each cluster. However, in case that this list is not enough, expert can browse on documents' title and abstract to help identify the most appropriate label for each cluster.

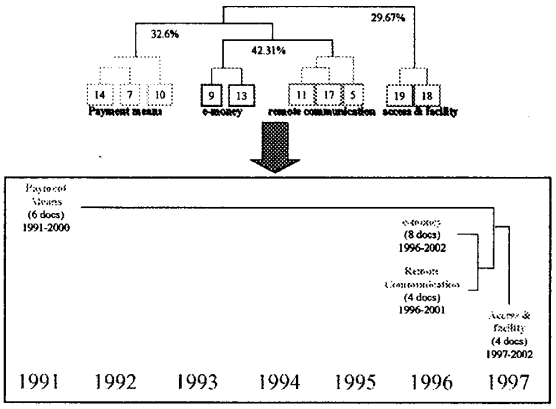

Figure 4.0 Visualizing Clustered Patent Documents

Lastly, visualize clustering results. This is a good representation of knowledge or information. Each cluster pertains to a particular technology or group, for example, cluster that talks about e-money and another cluster that talks about remote communication in payment. Each cluster may contain one or more documents closely related to each other and quite dissimilar to other cluster. From these documents in one cluster, one could get the earliest filing date wherein it could show that this cluster talks about this particular technology and the earliest filing date could be the start of it. Then, year of the earliest filing date could be used in mapping each cluster in time series fashion.

Also, each cluster could be explored to grasp more idea about a particular technology. In figure 5, from e-money cluster, when explored, one can see that "secure transmission" and "transferring of money" when merged is related to e-money holding device or somehow leads to development of such device. And obviously, there is technology on secure transmission or transmission itself before electronic funds transfer.

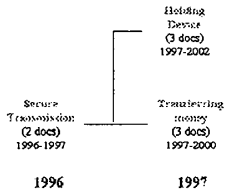

Figure 5.0 "e-money" cluster

IV Conclusion and Future work

This research had just attempt to use clustering techniques to cluster patent documents under same international patent classification, find its relationship and be able to visualize it in time series fashion to present knowledge or information. Patent documents itself are full of knowledge wherein one could learn and think of improvement or another invention related to it. So with analyzing patent documents collection and discovering knowledge from revealed clustered documents. And this work cannot be fully automated and would always require an expert to produce one patent map. However, improvements on natural language processing, labeling and testing this framework on larger data sets as well as developing program based on the proposed framework are potential future work.

[1] Wang Yong, Hodges Julia. "Document Clustering with Hierarchical Algorithm". http://www.cs.msstate.edu/~ywang/papers/ Accessed March 12, 2007
[2] Zhao Ying, Karypis George. "Hierarchical Clustering Algorithms for document datasets". Data Mining and Knowledge Discovery. Springer 2005. pp 141-168
[3] Weis, S., Indurkhya N., Zhang T., Damerau F. Text Mining Predictive Methods for Analyzing Unstructured Information. USA. Springer Science+Business Media, Inc. 2005.
[4] CLUTO: Clustering Toolkit. http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download
[5] USPTO Website. http://www.uspto.gov/patft/
[6] David G., Ophir F., Information Retrieval Algorithms and Heuristics 2nd Edition. Netherlands. Springer. 2004.
[7] Jong Hwan Suh, Sang Chan Park. "A New Visualization Method for Patent Map: Application to Ubiquitous Computing Technology". 2nd International Conference ADMA 2006 Proceedings. Page 566-573