A New Hierarchical Clustering Algorithm for Documents based on Tolerance Rough Set Model

Saori Kawasaki

School of Knowledge Science,
Japan Advanced Institute of Science and Technology
September 2000

Keywords: hierarchical clustering, text clustering, rough sets, tolerance rough set model, document similarity.

1. Objective

This research concerns with an extension of rough set theory known as tolerance rough set model (TRSM) and its application to text processing, especially it aims to develop a TRSM-based hierarchical clustering algorithm for documents. When dealing with a large set of documents, clustering is expected to provide users its structure that can use to improve efficiency and effectiveness in doing specified tasks. To achieve a TRSM-based hierarchical clustering with good quality, three issues are investigated within the framework of hierarchical agglomerative clustering: how to represent documents and clusters of documents, how to calculate distances between documents or clusters, how to reduce the computation costs in producing the hierarchy. This paper tries to answer those questions, to implement the algorithm and to do experiments in order to evaluate and validate the proposed algorithm.

2. Background

When handling a large set of objects, a common strategy is to divide the whole data set into subsets of related objects and to select appropriate parts for the processing purpose. Clustering is a powerful technique to achieve this aim by grouping objects into clusters using their similarities. When applying clustering to documents, the most crucial problem is how to consider the similarity between documents. Different similarity coefficients have been used in

the literature that often based on common terms included in documents, but only employing them cannot avoid making zero similarity between closed documents that do not share any common terms. Therefore, several methods using pre-defined dictionaries, term and phrases co-occurrence or patterns have been proposed for dealing with semantic relations of terms. One of them is based on theory of rough sets proposed by Pawlak in 1982. In the original rough set theory, a set is expressed by its lower and upper approximations regarding some equivalence relations. However, equivalence classes of terms are not suitable because of term ambiguities. A framework for generalized approximation spaces of rough set was formulated by Skowron allowing overlapping classes that require only two reflexive and symmetric properties. Inspired by the work of Skowron, Ho and Funakoshi recently developed an early version of tolerance rough set model for text processing.

3. A TRSM-based hierarchical clustering algorithm

Hierarchical clustering framework

There are two types of approaches to hierarchical clustering: Top-down and bottom-up. Since, it is important but too difficult to divide first several nodes from top-down, the major approach in hierarchical clustering is bottom-up clustering called HACM. The procedure outlines are as follows:

(1) assign all documents as clusters at leaf nodes
(2) calculate distances between all un-merged clusters
(3) assign a new cluster with paired clusters of maximum similarity
(4) repeat (2) and (3) until there is only one un-merged cluster.

Definition of tolerance spaces

Approximations of documents are essential in this algorithm. They are made of tolerance classes of terms. The tolerance class of a term consists of all terms those have the co-occurrence with this term in documents more than the given threshold.

Representation of documents

When the textual database consists of full texts and if not to use their grammatical structures, it is very redundant to use full text during distance calculation. To make compact sets of documents and reduce noise, extracted keywords can represent each document. The frequency of a term in a document and the frequency of documents including this term in the whole database allow us to estimate the importance of this term in the document. Each document can be represented by a set of terms weighted by their importance.

Distance between Documents

The distance between documents used in this work is cosine of term weight vector of

documents. For the cosine calculation, upper approximations of documents are employed to enrich the relation between documents and avoid zero similarity.

Representatives of clusters and distance between clusters

There are several links for measuring distance between clusters like single-link and complete-link. For the reasonable expression and computational complexity, methods of averaging distances between all documents included in clusters seem suitable for representing distance between clusters. Instead of it, the distance between representatives of clusters is used as the distance between clusters to reduce the computations. For enabling it, the representatives of a cluster are made from terms in documents included in the cluster with normalized weights. It is also employed here the quick sort algorithm for attempting to make a compact computation in having minimum distance among all document pairs.

4. Implementation and Experiments

Implementation of the algorithm

The whole procedure constructing a hierarchy form full text collection consists of two phases: the clustering itself and the preprocessing of full text. The preprocess includes (1) keywords extraction, (2) generation of tolerance classes of keywords, and (3) generation of upper and lower approximations of documents from tolerance classes of terms.

Test data and evaluation

The test collections MED, CISI, CRAN, CACM, JSAI are used to evaluate precision and recall of the method. As results, the TRSM cluster based retrieval achieved better precision than TRSM full search that achieved almost same result with exact match. An attempt has been done on clustering of Reuters collection that is typically used for evaluating text categorization. For the clustering validation, clustering stability and tendency are checked and the results indicate that the algorithm can provide a valid clustering for test collections. The computational complexities of this method reaches $O(N\log N)$ in keyword extraction, $O(N+M)$ for the tolerance spaces generation, and $O(M^2+N)$ for the hierarchical clustering.

5. Summary and further work

The proposed algorithm has interesting results in experiments, particularly for efficiency and precision though the advantage of recall is not improved. Some improvements can be done further such as using similarity relations which discharging the reflexive property from the tolerance relation to enrich the class expression of rough sets, or applying lower approximations to specified clusters. Also a combination of non-hierarchical clustering and hierarchical clustering seems to be a promising solution for very large databases.