Algorithms for Soft Document Clustering

Michal Rojček1, Igor Mokriš2

1Department of Informatics, Faculty of Pedagogic, Catholic University in Ružomberok, Hrabovská cesta 1, 03401 Ružomberok, Slovakia
   michal.rojcek@ku.sk

2Institute of Informatics, Slovak Academy of Sciences, Dúbravská cesta 9, 845 07 Bratislava 45, Slovakia
   igor.mokris@savba.sk

Abstract. Aim of this paper is to highlight the possibilities of clustering algorithms called as the "soft clustering" algorithms. The traditional approach "hard clustering" allows us to include only one document cluster. Soft clustering algorithms, like Fuzzy C-means (FCM), Word Base Soft Clustering (WBSC) Similarity-Based Soft Clustering Algorithm (SISC) and Kondadadi and Kozma modified ART (KMART), allow to aggregate similar documents to more than one cluster with different level of competence and therefore better reflect the nature and classification of the document. We anticipate that this approach is well suited for addressing homonyms. The contribution is devoted to compare different clustering algorithms due to their ability to be inserted into several clusters of documents.

1 Introduction

On the Web today we are faced with a multitude of different documents. We are not so interested if the search engine finds some results, but whether the results will be relevant for some of us. One of the following searched text document contains very often several different or similar topics. This overlap of topics can cause erroneous classification of the document. Many clustering algorithms such as "hard clustering" assigns each document to exactly one cluster. Then the user has difficulties to discover such information, which overlap each other in a variety of topics. For example, a document on common European bonds could fall as to the economy category, as well as to the category of policy.

Clustering, which allows to include the document to more categories called "soft clustering". This means that each document can belong to multiple clusters, and there rate to determine the association between each and every cluster of documents. It has the following advantages:

- a document can belong to multiple clusters, and thus can be discovered, several themes in the document,
- there may be clusters that contain a combination of subjects. For example, in experiments with Kmart in the system [1] has used a set of documents relating to
baseball, movies, and movies about baseball. Kmart has created three clusters of documents about baseball, movies and movies about baseball. Hard clustering algorithms (such as K-means) failed to create a cluster of baseball films,

- the rate of association between clusters and documents can be used as a measure of relevance to the order of suitability documents.

1. Key Requirements for Web Document Clustering

As described in [4] the following are the key requirements for web document clustering methods:
1. relevance: the method ought to produce that group documents relevant to the user’s query,
2. browsable summaries: the user needs to determine at a glance whether a cluster’s contents are of interest. Ranked lists of the cluster may infact difficult to browse. Therefore the method has to provide concise and accurate descriptions of the clusters,
3. overlap: since documents have multiple topics, it is important to avoid confining each document to only one cluster,
4. snippet – tolerance: the method ought to produce high quality cluster even when it only has access to the snippets returned by the search engines, as most users are unwilling to wait while the system downloads the original documents off the web,
5. speed: a very patient user might sift through 100 documents in a ranked list presentation. Clustering on the other hand allows user to browse several related documents. Therefore the clustering method ought to be able to cluster up to one thousand snippets in a few seconds. For the impatient user, each second counts,
6. incrementality: to save time, the method should start to process each snippet as soon as it is received over the web.

2. Fuzzy C-means (FCM)

Many soft clustering algorithms employ the idea of fuzziness in their methods. One of the most common fuzzy clustering algorithms is Fuzzy C-means. It was first reported by Dunn in 1972. FCM is based on the Partition clustering algorithm, iterating over the data sets until the values of the membership function stabilizes. FCM has been used in many applications like medical diagnosis, image analysis, irrigation design and automatic target recognition. Other fuzzy algorithm techniques such as Self-Organizing Maps, also abounds [1].

One drawback of fuzzy algorithms is that they are slow compared to non-fuzzy algorithms. Fuzzy clustering algorithms tend to be iterative, and typical fuzzy clustering algorithms require repeatedly calculating the associations between every cluster/document pair.
3. Word Base Soft Clustering (WBSC)

WBSC approach was proposed by King – Ip Lin, Ravikumar [2]. It first forms initial clusters of the documents, with each cluster representing a single word for instance, WBSC forms a cluster for the word ‘tiger’ made up of all the documents that contain the word ‘tiger’. After that, WBSC merges similar clusters – Clusters are similar if they contain the similar set of documents – using a hierarchical based approach until some stopping criterion is reached. At the end, the clusters are displayed based on the words associated with them. It consists of 3 steps that is cluster initialization, Cluster building and Display the result [3].

4. Similarity Based Soft Clustering (SISC)

This algorithm is a soft clustering algorithm. It uses modified fuzzy C Means algorithm to cluster set of documents based on a given similarity measures. It use a randomization approach that enables it to avoid a lot of computation needed in a traditional fuzzy clustering algorithms. At each iteration, it computes a similarity measure between a cluster and a document with a probability proportional to the proximity of the similarity measure to the threshold measure [3].

SISC can broadly divided into four steps:
1. a preprocessing steps to cleanup and transfer a data,
2. a initial cluster generation steps to initialize the cluster,
3. an iterative step to build the cluster,
4. and post processing steps to present the result.

5. Kondadadi and Kozma Modified ART (KMAR)

To create clusters with Kmart algorithm using a modified version of Fuzzy ART network. In [1] it was proposed the replacement of existing Fuzzy ART algorithm in order to apply the Soft clustering. Instead of choosing a maximum similarity of the category and use the test vigilance (vigilance test) to check if the category is fairly close to the input pattern, it can be checked in each category layer F2 by applying a vigilance test. If the category meets the vigilance test, the input document is inserted into the particular category.

Similarity measurement is calculated in the test defining the degree of alertness of the input pattern belonging to the current cluster. This allows the document in several clusters with varying degrees of competence. All prototypes that pass the vigilance test are updated by the adaptation rules Fuzzy ART network.

This modification provides additional benefits resulting from the above-mentioned Soft clustering:

- fuzzy ART is usually time consuming because it requires iterative search to find the winning category that meets the vigilance test. In the described modification,
the search is not necessary because each F2 node is already checked. This makes this model computationally less demanding,

- another advantage is removed step of choosing categories, we avoid using the parameter $\beta$ choice (choice parameter). This will reduce the number of user-defined parameters in the system. This modification does not violate the basic principle of ART networks, i.e. to avoid the dilemma of stability and plasticity. Kmart is still agglutinative incremental algorithm and before learning from re-entry access control and input pattern will learn only if it meets one of the stored patterns in a certain tolerance.

The last step is to see in Kmart representative keywords for each cluster, formed in the previous phase. This allows users to distinguish between different clusters. For each cluster are ranked words according to the number of documents in the cluster in which the word appears and similarities of documents (defined wakefulness test) in which the word appears. Usually you first 7-10 words as keywords.

6. Conclusion

Many soft clustering algorithms implements the idea of fuzziness in their method. The main advantage of KMART over most of fuzzy clustering algorithm is that the number of clusters is decided dynamically.

References

Acknowledgement

This work was supported by the Slovak Science Agency VEGA No. 2/0211/09 and VEGA No. 2/0184/10.