Assessment of the Performance of Fuzzy Cluster Analysis in the Classification of RFC Documents

M. E. S. Mendes, L. Sacks
Dept. of Electronic and Electrical Engineering
University College London
{mmendes@ee.ucl.ac.uk}

Abstract: *This paper describes the use of fuzzy cluster analysis in the process of finding knowledge from meta-data. A collection of RFC documents has been used to analyse the behaviour of such technique on finding similarities and establishing groups among these documents. The fuzzy clustering resulted in a 85% match to our prior expectations.*

1 Introduction

The work described in this paper has been carried out in the context of a network-based teaching and learning system. In such systems, the need to find relevant learning resources to a particular knowledge domain arises. This need comes not only from a learning perspective, but also from an authoring perspective.

From the authoring point of view, the re-use of documents to facilitate the delivery of new courses is seen as an important advantage. This helps to minimise the authoring task, as long as there is a way of locating relevant learning resources for the course in question.

From the learning perspective, we must keep in mind that students have different profiles in terms of background knowledge, learning objectives, preferred learning styles, etc. Despite the fact that online courses might present a well-defined structure, which is conceived by the teacher, such structure should not be seen as a rigid track to follow, but as an orientation track. To cope with the different student profiles, the system should allow an adaptive navigation of the document space, based on a relevance measure of other documents to the ones that are part of the defined tracks.

The question that arises is how to measure relevance and how to organise documents in an abstract knowledge space. To evaluate how the knowledge space could be formed, an experiment has been carried out with a collection of RFC documents. As a candidate technology, the use of fuzzy cluster analysis has been explored.

2 Fuzzy Clustering

Clustering algorithms are used to find groups in unlabeled data, based on a similarity measure between the data patterns (elements). This means that similar patterns are placed together in the same cluster.

The main difference between fuzzy clustering and other clustering techniques is that it generates fuzzy partitions of the data instead of hard partitions. Therefore, data patterns may belong to several clusters, having in each cluster different membership values.

Thinking on the on-line repository of educational documents, this membership concept becomes quite important. It means that there is a way to situate documents in several knowledge domains, with different weights in each of them.

The *fuzzy c-means algorithm* [1] was the technique chosen to conduct our experiment with the RFC documents and therefore, a description on it follows.

Fuzzy c-means algorithm

Let $X \equiv [x_1, x_2, \ldots, x_N]$ be $N \times d$ matrix where N is the number of patterns (elements) and d is the dimensionality of the patterns (number of features).

Let U be the universe of all possible partitions of X and c an integer $1 < c < N$, representing the number of clusters. A partition $U \equiv [u_{ai}] \in U$ is a $c \times N$ matrix that satisfies the following three conditions:

1. $u_{ai} \in [0,1], \forall a = 1, \ldots, c, \forall i$ 2. $\sum_{a=1}^{c} u_{ai} = 1, \forall i$ 3. $0 < \sum_{i=1}^{N} u_{ai} < N, \forall a = 1, \ldots, c$

Let $V \equiv [v_1, v_2, \ldots, v_c]$ be a $c \times d$ matrix representing the centres of the clusters and $\|\cdot\|$ a distance function, which is the measure of similarity between patterns. The FCM goal is to minimize an objective function J_m, which is a weighted sum of squared errors:

$$J_m(U,V) = \sum_{i=1}^{N}\sum_{a=i}^{c} {u_{ai}}^m \left\| x_i - v_a \right\|^2 ,$$

x_i – i^{th} pattern of X
v_a – centre (prototype) of the a^{th} cluster
u_{ai} – grade of membership of x_i in cluster a
m – fuzzification parameter ($m>1$)

Step1. Select the number of clusters c, the termination criteria $e > 0$, the value for $m > 1$ and initialise the partition matrix $U^{(0)} \in U$.

Step2. Compute the prototypes of the clusters and update the partition $U^{(t)}$ according to:

$$v_a^{(t)} = \frac{\sum_{i=1}^{N} \left(u_{ai}^{(t-1)}\right)^m \cdot x_i}{\sum_{i=1}^{N} \left(u_{ai}^{(t-1)}\right)^m} , \; a = 1, 2, \ldots, c \qquad u_{ai}^{(t)} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\left\| x_i - v_a^{(t)} \right\|}{\left\| x_i - v_j^{(t)} \right\|} \right)^{2/(m-1)}} , \; \begin{matrix} a = 1, 2, \ldots, c \\ i = 1, 2, \ldots, N \end{matrix}$$

Step3. If the termination criteria is satisfied, $\left\| v^{(t+1)} - v^{(t)} \right\| < e$, stop. Else, go to step 2.

3 The process of finding similarities among RFC documents

In order to analyse whether fuzzy clustering would be suitable to form the weighted knowledge space, we applied this technique to a sample collection of documents. We chose from the many RFC documents that are available on the Internet, a subset containing the ones that described standards. Along with the documents, we were able to find meta-data [2] about them, which was basically a set of indexing terms for each RFC. The experiment carried out followed the process represented in Figure 1.

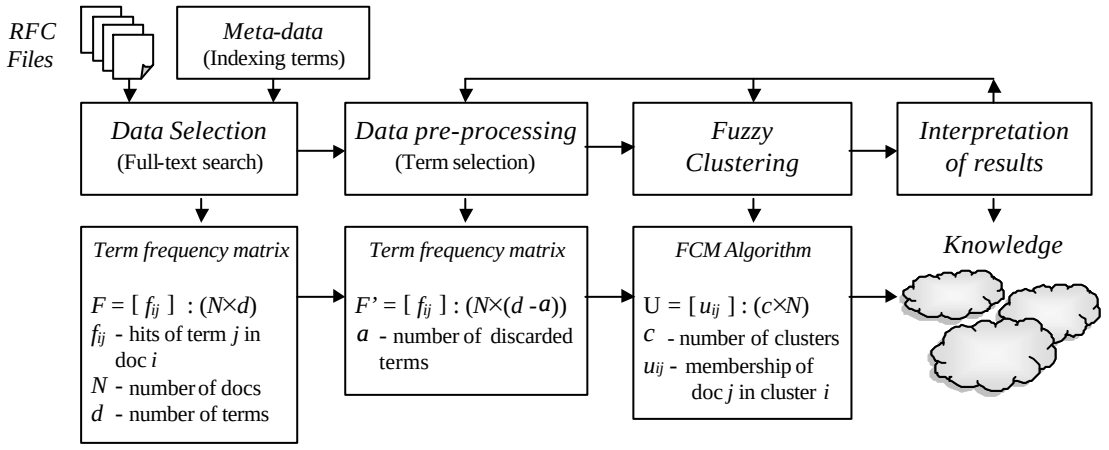


Figure 1 – The process of finding knowledge from meta-data

Assuming as *meta-data* the available set of indexing terms, the first step consisted on a *full-text search* of the RFC files to obtain the term frequency information. Then followed the *pre-processing* phase, which aimed at selecting only the relevant terms for the

clustering process. There are several text indexing techniques referred in literature [3][4] that allow to determine the significance of a given term based on its document frequency information. One measure of significance is the *term specificity* (1). Another measure is based on Information Theory, considering that a term carries more information when it has a lower e*ntropy* value (2).

$$sp = \log \frac{N}{n_j} \quad (1) \qquad H = \sum_{i=1}^{N} p_i \cdot \log_2 \frac{1}{p_i}, \quad \text{with } p_i = \frac{f_{ij}}{F_j} \text{ and } F_j = \sum_{i=1}^{N} f_{ij} \quad (2)$$

(N = total number of documents, n_j = number of documents that contain the term j)

The *specificity* measure is useful for identifying terms that are too specific, which means that they only appear in a very small percentage of documents and consequently they might be irrelevant for the clustering. As for the *entropy* measure, terms that exhibit high entropy values might be seen as noise in the clustering process and so they should be discarded. The maximum value for the *entropy* occurs when a term has equal probability in every document:

$$H_{\max} = \sum_{i=1}^{N} \frac{f_j}{N \cdot f_j} \cdot \log_2 \frac{N \cdot f_j}{f_j} = \log_2 N \quad (3)$$

In our experiment we were mainly focused on the *fuzzy clustering* phase, leaving to a second stage issues like:

- determining the best indexing terms for each document using an automatic indexing algorithm (and then compare those terms with the ones that were currently used);
- analysing to what extent does the *pre-processing* phase affect the clustering results, in order to determine how robust the fuzzy clustering is to the inclusion of "noisy" terms.

It was necessary to select a distance function to apply in the FCM algorithm. We tried two different functions, one was the Euclidean distance (4) and the other was a distance based on the similarity function proposed in [5]. This similarity function (5) represents the proximity between two document vectors D_a and D_b, based on the weights each term has in each of them.

$$dist(x_i, v_a) = \sqrt{\sum_{j=1}^{d} \left(x_{i,j} - v_{a,j}\right)^2} = \left\| x_i - v_a \right\| \quad (4)$$

$$sim(D_a, D_b) = \sum_{j=1}^{d} w_{a,j} \cdot w_{b,j}, \qquad \text{where } w_{i,j} = \frac{f_{ij} \cdot sp_j}{\sqrt{\sum_{k=1}^{d} \left(f_{ik} \cdot sp_k\right)^2}} \quad (5)$$

Since similarity is inversely proportional to distance we derived the following expression and applied it in the FCM algorithm:

$$dist(x_i, v_a) = \frac{1}{sim(x_i, v_a)} = \frac{1}{\sum_{j=1}^{d} w_{i,j} \cdot w_{a,j}} = \left\| x_i - v_a \right\| \quad (6)$$

The last phase consisted in the *analysis* of the clustering results. Since our objective was to evaluate the performance of the fuzzy clustering, we had to come up with a reference clustering for comparing the results. Based on what we knew about the RFC documents, we were able produce such reference.

4 Results

As the input for the knowledge discovery process we had a sample document collection, composed by 73 RFCs, and a list of 81 indexing terms.

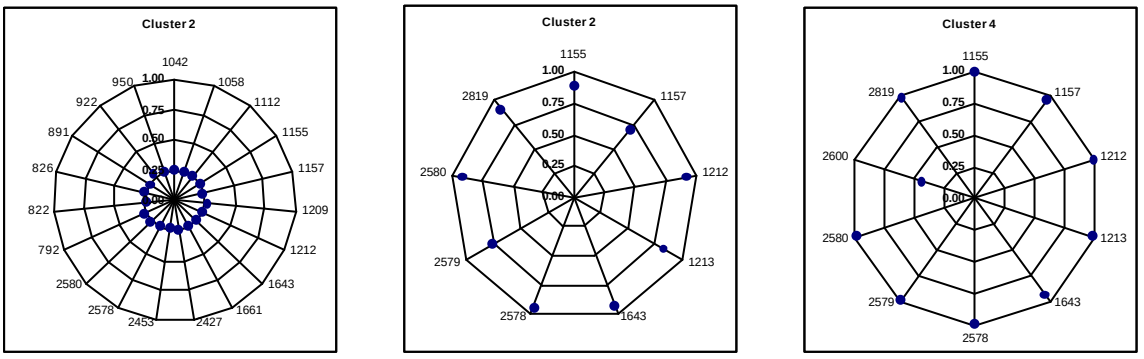
Even though we were not concerned at this stage with the pre-processing phase, we did some tests using a combination of *specificity* and *entropy* measures in order to see which terms would be filtered. As an example, when we defined a threshold for the *noise* values to be below 75%×H_{max} the following terms were filtered: “address”, “data”, “first”, “information”, “internet”, “level”, “local”, “network”, “protocol” and “time”. This filtering seems to make sense, since these words are likely to appear in almost every RFC and therefore, they wouldn’t improve the clustering results.

Having performed several experiments with the FCM algorithm using the two distance functions presented in section 3, we observed that the similarity-based distance function (eq.6) produced much better results than the Euclidean distance (eq.5). When pre-processing was applied (without being too restrictive with the filtering thresholds), we observed that the results were slightly improved. Table 1 summarises the performance of the fuzzy clustering for 5 clusters, which was calculated by comparing the results with a reference clustering. We can see that the best results matched our expectations in 85%.

Euclidean distance function		Similarity-based distance function	
Without pre-processing	With pre-processing	Without pre-processing	With pre-processing
61%	62%	82%	**85%**

Table 1 – Performance of the fuzzy clustering for c = 5 clusters

When the Euclidean distance was used, we had to perform several iterations of the FCM algorithm in order to find 5 clusters, whereas with the similarity-based distance only one iteration was needed. The plots presented in Figure 2, show one of the five clusters obtained. The values around the plots represent RFC numbers and the vertical axes represent the degree of membership.



a) **b)** **c)**

Figure 2 – Clustering of RFCs that are related with “network management” using: a) Euclidean distance without pre-processing, b) Similarity-based distance without pre-processing, c) Similarity-based distance with pre-processing

We can observe that in cases b) and c) we were able to isolate RFCs that were related with “Network Management”. In case a), only 67% of those RFCs were grouped in the same cluster. We can also see that the membership values in case a) are quite low (around 0.25). As for the other two cases, c) produced higher membership values than b). Another observation is that RFC 2600 is present in c), with a membership below 0.50, but not in b). This is the only RFC in that cluster that is not clearly related with “Network Management”.

5 Conclusions

The fuzzy clustering technique proved to be quite effective in grouping the RFC documents based on their resemblance. One conclusion that comes out of our experiment is that the distance function derived from the similarity expression produces very good results, even without pre-processing the data. Some issues remain for further investigation:

- having observed the positive effect of the pre-processing phase in the clustering results, we should explore this phase in more detail to try to maximise the clustering performance
- the impact of removing documents or adding new documents to the original collection should be evaluated.

Acknowledgment

This work has been supported by *Fundação para a Ciência e a Tecnologia* through the PRAXIS XXI scholarship programme.

References

[1] J.C. Bezdek. “Pattern Recognition with Fuzzy Objective Function Algorithms”. Plenum Press, New York, 1981.

[2] http://www.garlic.com/~lynn/rfckeyw.htm

[3] C. Faloutsos, D. W. Oard. “A survey of information retrieval and filtering methods”. Technical Report CS-TR-3514, University of Maryland, 1995. (http://www.enee.umd.edu/medlab/filter/papers/survey.ps)

[4] G. Salton. “A Theory of Indexing”. Philadelphia: Society for Industrial and Applied Mathematics , 1975.

[5] G. Slaton, J. Allan, C. Buckley. “Automatic structuring and retrieval of large text files”. Communications of the ACM, Vol. 37, No. 2 (Feb. 1994), pp. 97-108