From Metadata to Fuzzy Knowledge Representation

M. E. S. Mendes, L. Sacks
Dept. of E&EE, University College London

Abstract: In this paper we present an approach to organize Web-accessible courseware according to knowledge domains, by means of fuzzy clustering. A new modified version of the Fuzzy C-Means clustering algorithm that employs a non-Euclidean metric is proposed and some preliminary trials with the new metric are presented. The experimental results show that the modified algorithm performs better than the original and that it fulfils the requirements of our Internet-based teaching and learning application.

1. Introduction

This paper reports on a knowledge representation framework for Internet-based teaching and learning applications. Research and development of one such application is being carried out by the CANDLE project (Collaborative And Network Distributed Learning Environment – http://www.candle.eu.org/). This project runs under the European Commission IST fifth framework programme and its main focus is on the delivery of courseware for the telematics domain over the Internet. CANDLE puts great emphasis on the creation of sharable and reusable teaching material to be used by a network of people from different universities and corporations all over Europe. The project is also concerned with increasing the flexibility of the learning process, to accommodate various pedagogical approaches and flexible usages of the courseware by learners.

Following the current paradigm of semantic interoperability among networked information resources, every learning material within CANDLE's repository is described by metadata and represented in XML (eXtensible Markup Language). The fundamental role of metadata is to enhance the process of information retrieval, by providing rich and machine "understandable" representations. The project is developing its own educational metadata scheme, which specializes IEEE's LOM (Learning Object Metadata) specification [1]. Although metadata provides very useful information for basic search and retrieval tools, in applications like CANDLE a more elaborate context-dependent retrieval mechanism is required. Learning objectives, pedagogical approaches and user profile are examples of variables that need to be considered for defining the appropriate set of links to relevant documents. The definition of relevance depends firstly on the set of subjects associated with each document and secondly on the usage context. This implies that a classification of documents in terms of knowledge domains needs to be provided and that a flexible knowledge representation framework needs to be developed for supporting different pedagogical models and learning styles.

CANDLE's metadata scheme includes a category to classify the courseware, based on its *knowledge space* location. This is done through the selection of the appropriate key words extracted from a predefined taxonomy of the telematics domain. The *knowledge space* can be built in several ways; our approach is to use fuzzy clustering to identify fuzzy relationships between learning materials and to dynamically discover the underlying knowledge structure using the metadata information. In the remaining sections of this paper we detail on this approach: some background and a new modified version of a well-known fuzzy clustering algorithm along with some experimental results are presented.

2. Fuzzy Clustering for Knowledge Representation

Our objective is to discover knowledge-based relations that may exist between learning material, based on their metadata descriptions. This can be seen as a document clustering problem whereby a suitable algorithm may be applied for organizing the XML documents and for discovering hidden or unobvious relations between them. Regardless of the algorithm or the data set, every clustering method aims at grouping data elements according to some similarity measure so that related elements are placed in the same cluster. For our application we find that fuzzy clustering techniques are more suitable than hard clustering methods like agglomerative hierarchical clustering [2] and the partitional K-Means, which are perhaps the most popular ones for document clustering. These methods generate hard clusters in the sense that each document is assigned to one and only one cluster. But in our case several sources of uncertainty can be identified, which need to be handled properly:

(a) Firstly one cannot expect the *knowledge space* representation to be completely accurate, because knowledge is abstract by nature and even experts might disagree on the correct knowledge definition. Thus, fuzzy document relations are more likely to represent the "true" knowledge structure than hard ones.

(b) Secondly the metadata concerning the classification of courseware represents the author's best attempt to define the subjects associated each material. But the tagging process is intrinsically imprecise since it reflects the author's subjective opinion.

The theory of fuzzy sets [3] provides the mathematical means to deal with uncertainty and fuzzy clustering brings together the ability to find unobvious relations in data sets with the ability to cope with uncertainty. Thus, a method capable of generating fuzzy clusters seems to be more appropriate. For our trials we chose the Fuzzy C-Means [4], which is one of the most popular fuzzy clustering methods. It generalizes the hard K-Means, by producing a fuzzy partition of the data space as opposed to a hard one. We decided to use this algorithm in our experiments for its simplicity and for being the fuzzy extension of a common document clustering technique.

3. Background for the Fuzzy Clustering Trials

3.1 Metric Concepts

In order to apply a clustering algorithm to a document collection it is necessary to have suitable document representations. In the well-known Vector Space Model (VSM) of information retrieval [5] each document is represented by a set of indexing terms in the form of a k-dimensional vector (3.1), where k is the total number of terms and w_{ij} represents the weight (or significance) of term j in document x_i. A term weighting system that has proved to perform well [6] – see equation (3.2) – considers the frequency of term j in document i - f_{ij}, the inverse document frequency (n_j = number of documents that contain term j and N = total number of documents) and a factor of document length normalization.

$$x_i = [w_{i1}\ w_{i2} \ldots w_{ik}] \qquad (3.1) \qquad\qquad w_{ij} = f_{ij}\cdot\log(N/n_j)\cdot\left[\sum_{t=1}^{k}\left(f_{it}\cdot\log(N/n_t)\right)^2\right]^{-1/2} \qquad (3.2)$$

A similar vector representation can be obtained for CANDLE's learning materials from their metadata descriptions. In this case, authors assign manually the indexing terms (key words) and the associated weights.

The weighted document vectors are suitable to be processed by clustering algorithms, which group documents according to their similarities. So, it is necessary to choose an appropriate similarity measure for the document space, like the function defined in (3.3). This metric is used in the VSM to compare document vectors with query vectors [6]. Since x_a and x_b are normalized weighted vectors the similarity function exhibits properties (3.4) and (3.5).

$$S(x_\alpha, x_\beta) = \sum_{j=1}^{k} w_{\alpha j}\cdot w_{\beta j} = x_\alpha \cdot x_\beta^{\,T} \quad (3.3) \qquad 0 \le S(x_\alpha, x_\beta) \le 1, \forall_{\alpha,\beta} \quad (3.4) \qquad S(x_\alpha, x_\alpha) = 1, \forall_\alpha \quad (3.5)$$

Some clustering algorithms group data elements based on dissimilarities or distances. A dissimilarity function can be obtained from the similarity metric defined in (3.3) by an appropriate transformation:

$$D(x_\alpha, x_\beta) = 1 - S(x_\alpha, x_\beta) = 1 - \sum_{j=1}^{k} w_{\alpha j}\cdot w_{\beta j} = 1 - x_\alpha \cdot x_\beta^{\,T} \qquad (3.6)$$

3.2 Fuzzy C-Means Algorithm

The Fuzzy C-Means algorithm (FCM) [4] is overviewed next. Given a data set with N elements each represented by k-dimensional feature vector, the FCM takes as input a ($N\times k$) matrix X=[x_i]. The number of clusters c ($1<c<N$) and the fuzzification parameter m ($m>1$) need to be selected initially. Also, a distance function $\|\cdot\|$ needs to be chosen, the most common being the Euclidean norm (3.7). The FCM runs iteratively to obtain the cluster centres – V=[v_α]: ($c\times k$) – and a partition matrix – U=[$u_{\alpha i}$]: ($c\times N$) – which contains the membership of each data element in each of the c clusters. Both the cluster centers and the partition matrix are computed optimizing the objective function defined in (3.8).

$$d_{i\alpha}^{\,2} = \left\|x_i - v_\alpha\right\|^2 = \sum_{j=1}^{k}(x_{ij} - v_{\alpha j})^2 \qquad (3.7) \qquad\qquad J_m(U,V) = \sum_{i=1}^{N}\sum_{\alpha=1}^{c} u_{\alpha i}^{\,m} d_{i\alpha}^{\,2} = \sum_{i=1}^{N}\sum_{\alpha=1}^{c} u_{\alpha i}^{\,m}\left\|x_i - v_\alpha\right\|^2 \qquad (3.8)$$

The FCM algorithm starts with a random initialisation of the partition matrix subject to the following three constraints:

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

At each iteration, the cluster centres and the grades of membership are updated according to (3.9) and (3.10) respectively. The algorithm ends when a termination criterion is met or the maximum number of iterations is achieved.

$$v_\alpha = \frac{\sum_{i=1}^{N} u_{\alpha i}^{\,m}\cdot x_i}{\sum_{i=1}^{N} u_{\alpha i}^{\,m}} \qquad (3.9) \qquad\qquad u_{\alpha i} = \left[\sum_{\beta=1}^{c}\left(\frac{d_{i\alpha}^{\,2}}{d_{i\beta}^{\,2}}\right)^{1/(m-1)}\right]^{-1} = \left[\sum_{\beta=1}^{c}\left(\frac{\left\|x_i - v_\alpha\right\|}{\left\|x_i - v_\beta\right\|}\right)^{2/(m-1)}\right]^{-1} \qquad (3.10)$$

3.3 Modified Fuzzy C-Means Algorithm

The Euclidean norm, which is frequently applied in the FCM algorithm, is not the most suitable for comparing document vectors, because the non-occurrence of the same terms in both documents is treated in the same way as the co-occurrence of words [7]. A suitable dissimilarity function for document vectors was introduced in (3.6). We decided to apply this metric for clustering documents, using the FCM approach. The modified objective function is similar to (3.8), but now the norm $\|\cdot\|^2$ is replaced by the function defined in (3.6):

$$J_m(U,V)=\sum_{i=1}^{N}\sum_{\alpha=1}^{c}u_{\alpha i}{}^{m}D_{i\alpha}=\sum_{i=1}^{N}\sum_{\alpha=1}^{c}u_{\alpha i}{}^{m}(1-\sum_{j=1}^{k}x_{ij}\cdot v_{\alpha j}) \qquad (3.11)$$

As the expression used to update of the clusters centres (3.9) was obtained considering the Euclidean distance we had to derive a new expression to work with the dissimilarity function – see equation (3.12). Details on this new development are presented in [7]. It can be proved that minimizing (3.11) with respect to u_{ai} leads to a similar result as in (3.10), but now $d_{ia}{}^2$ and $d_{ib}{}^2$ are replaced by D_{ia} and D_{ib},. The expression for u_{ai} is shown in (3.13).

$$v_\alpha=\sum_{i=1}^{N}u_{\alpha i}{}^{m}x_i\cdot\left[\sum_{j=1}^{k}\left(\sum_{i=1}^{N}u_{\alpha i}{}^{m}x_{ij}\right)^2\right]^{-1/2} \qquad (3.12)$$

$$u_{\alpha i}=\left[\sum_{\beta=1}^{c}\left(\frac{D_{i\alpha}}{D_{i\beta}}\right)^{1/(m-1)}\right]^{-1}=\left[\sum_{\beta=1}^{c}\left(\frac{1-x_i\cdot v_\alpha{}^T}{1-x_i\cdot v_\beta{}^T}\right)^{1/(m-1)}\right]^{-1} \qquad (3.13)$$

The new modified FCM runs similarly to the original FCM, differing only on the expressions used to update v_a and to calculate the distances.

3.4 Fuzziness of the Document Clusters

It is known that increasing values of m lead to a fuzzier partition matrix. For the reasons presented in section 2, the more fuzzy the results, the more flexible will be the use of the discovered document relations. However, there needs to be a compromise between the amount of fuzziness and capability to obtain good clusters and reason from those relations. If all documents end up with the same membership in every cluster, the conclusion will be that they are all equally related to each other. A simple cluster validity measure that indicates the closeness of a fuzzy partition to a hard one is the Partition Entropy (*PE*) [4]:

$$PE=-\frac{1}{N}\sum_{i=1}^{N}\sum_{\alpha=1}^{c}u_{\alpha i}\log_a(u_{\alpha i}) \qquad (3.14)$$

The possible values of *PE* range from 0 – when U is *hard* – to $\log_a(c)$ – when every data element has equal membership in every cluster ($u_{ai}=1/c$).

4. Fuzzy Clustering Trials

The aim of our experiments was to investigate whether or not fuzzy clustering was suitable for our purposes. We carried out several trials to assess and compare the performance of the FCM applying different metric concepts: the Euclidean distance and the dissimilarity function. This section reports on our experiments.

4.1 Data Set Description

The process of populating CANDLE's database with learning materials has just recently started. As we had to simulate CANDLE database, we decided to work with a familiar collection of text document. We selected a set of RFC text documents (that describe standard protocols and policies of the Internet). Each of the documents was automatically indexed with keywords from an existing taxonomy [8]. Document vectors as in (3.1) were generated and organized as rows of a ($N\times k$) matrix, where N=67 was the collection size and k=465 was the total number of indexing terms. We manually created a clustering benchmark based on our knowledge of the documents' contents, complemented by the indexing information found in [8]. The benchmark indicated that the RFCs could be distributed into 6 fairly homogeneous clusters although some of the documents could have been attributed to more than one cluster [9].

4.2 Experimental Results

We performed several trials with the FCM algorithm applying both the Euclidean distance (FCM-ED) and the dissimilarity function (FCM-DF). For each case we fixed the convergence threshold to 10^{-4} and the maximum number of iterations to 300. For the FCM-DF trials we created a document matrix of term weights ($X_1=[w_{ij}]$) using (3.2). For the FCM-ED trials we also generated a matrix of term frequencies ($X_2=[f_{ij}]$).

Trial 1: In this trial the objective was to analyse whether the FCM algorithm would be able to generate a good partition of the document collection. We fixed ***c*=6** (as our benchmark indicated) and we set ***m*=1.1** so that the clusters would be close to the non-fuzzy case. To compare the results with the reference clustering we applied the maximum membership criterion to generate hard clusters from the fuzzy ones. We ran the FCM-ED using as input $\mathbf{X_1}$ and also $\mathbf{X_2}$, and we ran the FCM-DF using as input $\mathbf{X_1}$.

We found out that the FCM-ED performed poorly when X_2 was used as input: even for such a low value of m, ~86% of the documents ended up in the same cluster. But when normalized weighted vectors were used both FCM-ED and FCM-DF generated clusters with a high degree of match with the reference.

Trial 2: The second trial was to compare the performance of the two metrics for several values the fuzzification parameter, $m\in[1.1,2.5]$, keeping ***c*=6**. We observed that with the Euclidean distance (FCM-ED using $\mathbf{X_2}$) the execution times of the algorithm increased exponentially as m increased – see Figure 1. We can see that these times are much higher than the ones obtained for the dissimilarity function (FCM-DF with $\mathbf{X_1}$). We verified that the CPU time required per iteration of the FCM algorithm was approximately the same regardless of the metric used. Thus, the increase of the total execution time is due to an increase of the number of iterations required until the convergence threshold is achieved. We can see on the plot that for $m \geq 1.8$ the FCM-ED is not able to converge in less than 300 iterations.

When $\mathbf{X}_1$ was used as input both for FCM-ED and FCM-DF, we verified that with the Euclidean distance the algorithm only generated good partitions for low values of m (≤ 1.3), whereas with the dissimilarity function higher degrees of fuzziness were acceptable without compromising the quality of the clusters [7].

Trial 3: The objective of this trial was again to compare the performance of the two metrics, but now for increasing number of clusters, $c \in [2,10]$, with m fixed at different values. We found out that regardless of the number of clusters the FCM-ED (with X_1) produced partitions with maximum fuzziness when m was higher than 1.3. This was not the case with the FCM-DF. To exemplify, Figure 2 presents the partition entropy (3.14) obtained with both metrics (for m=1.5). The plot shows that with the Euclidean distance the partition entropy is always maximal even for few clusters, which is not the case with the dissimilarity function that produced good fuzzy partitions for increasing number of clusters. When c was higher that 6 we analysed the contents of the clusters and compared them with the benchmark. We discovered that there was still a high degree of match and more importantly, we found out that the algorithm was successfully able to identify good sub-clusters within the reference ones.

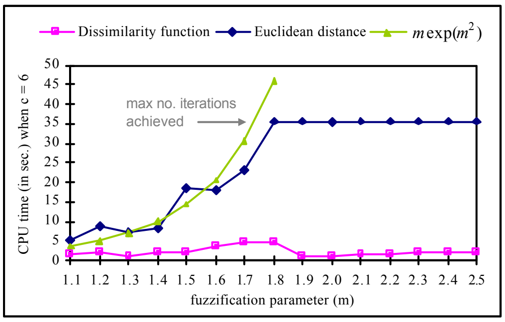

Figure 1. Execution times of the FCM-DF using X_1 and the FCM-ED using X_2, for increasing values of m (c=6 clusters)

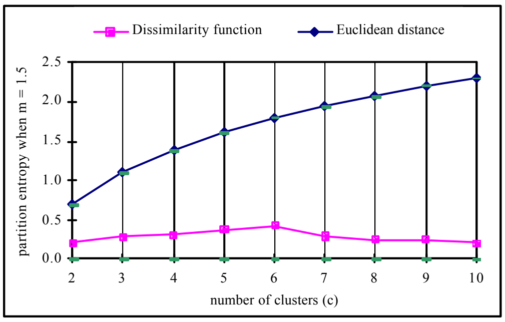

Figure 2. Partition entropy of the FCM-DF and the FCM-ED (both using X_1) for increasing values c (m = 1.5)

5. Conclusions and Further Work

In this paper, we presented an approach to represent knowledge domains through the discovery of fuzzy relationships between leaning materials. We presented a new modified version of the Fuzzy C-Means clustering algorithm that employed a dissimilarity function common in traditional information retrieval systems. Our experiments with the RFC document collection showed that the FCM algorithm produces poor results for term frequency vectors, but when normalized weighted vectors are used the FCM successfully approximates the reference clusters. We also verified that with the Euclidean distance good partitions were generated, but only for low values of m, whereas with the dissimilarity function good clusters were obtained for higher degrees of fuzziness. This is an important result, which indicates that the FCM with the new metric fulfils our requirements regarding knowledge-based organization of CANDLE's learning materials. Another important conclusion is that the algorithm is capable of identifying sub-structures within the clusters, which indicates that a hierarchical organisation of the learning materials is possible through a nested refinement of the fuzzy partitions. In the near future, our investigation will address this issue and also issues regarding the incremental update of the fuzzy clusters to deal with new document arrivals in the database.

6. Acknowledgments

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

7. References

[1] IEEE LOM Working Group. "Draft Standard for Learning Object Metadata," Nov. 2000.

[2] P. Willett. "Recent trends in hierarchical document clustering: a critical review," Information Processing and Management, Vol. 24, No. 5, 1988, pp. 577-597.

[3] L. A. Zadeh. "Fuzzy Sets," Information and Control, Vol. 8, 1965, pp. 338-353.

[4] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[5] G. Salton, J. M. McGill. *Introduction to modern information retrieval*. McGraw-Hill, New York, 1983.

[6] G. Salton, 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.

[7] M. E. S. Mendes, L. Sacks. "Dynamic Knowledge Representation for E-Learning Applications." To appear in Proc. of the 2001 BISC International Workshop on Fuzzy Logic and the Internet, FLINT 2001, University of California Berkeley, Aug. 2001.

[8] L. Wheeler. *IETF RFC Index*. Available at: http://www.garlic.com/~lynn/rfcietf.htm

[9] M. E. S. Mendes, L. Sacks. "Assessment of the Performance of Fuzzy Cluster Analysis in the Classification of RFC Documents," Proc. of the London Communications Symposium, Sep. 2000, London.