Dynamic Knowledge Representation for E-Learning Applications

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

Abstract
In this paper, we present an approach to organize E-Learning materials according to knowledge domains, by means of fuzzy clustering analysis. A new modified version of the Fuzzy C-Means clustering algorithm has been derived to employ a non-Euclidean metric, which is common in traditional information retrieval systems. The preliminary trials with this modified algorithm show that it performs better than the original one, when used for clustering text documents.

1. Introduction

Information resources play an important role in almost every Internet service, and especially in E-Learning (Internet-based teaching and learning). Functionalities like information search and retrieval are a must. But, in the case of E-Learning, such functionalities need to be complemented by taking into account issues like, for instance, learning objectives, pedagogical approaches and learner profile. Thus, there is a context for retrieving information. It is necessary to define which learning materials are relevant to a particular user, who wants to learn about a particular subject.

To determine which of the materials available in the courseware database are relevant, two components should be considered. Firstly, the set of subjects associated to each material should be identified. Thus, there needs to be a way to classify and organize materials in terms of knowledge domains. Secondly, the learning context should be identified, since the learning goals and pedagogical models may impact on the way materials are structured and consequently, on the definition of relevant links.

In this paper we present our approach to obtain the first input for the computation of relevance. The subsequent sections are organized as follows. The CANDLE project is briefly overviewed in section 2. In section 3 issues associated with the representation and retrieval of learning materials in the project's context are presented. The argument for employing fuzzy clustering to organize learning materials and to build a domain knowledge representation is presented in section 4. In section 5, we present the background for our experiments with fuzzy clustering applied to text documents, which are then detailed in section 6. In section 7 we present our conclusions and further research issues.

2. The CANDLE Project

Presently, there are many initiatives around the world working on the development of Internet-based teaching and learning applications, both for education and training purposes. One of those initiatives is the CANDLE[1] project, which is a European Commission shared cost research project under the IST fifth framework programme. Its main focus is on the delivery of courseware, for the telematics domain, over the Internet.
The project's major long-term objectives focus on: -

(a) helping educators to be increasingly successful;
(b) increasing the flexibility of the learning process, to accommodate various pedagogical approaches and flexible usages of the learning materials;
(c) developing a suitable technical framework to allow learning materials to be sharable and reusable, enabling the rapid development and deployment of new courseware.

So, CANDLE puts a strong emphasis on the learning materials, both in terms of representation and retrieval, for allowing the flexibility and reusability required.

3. Representation and Retrieval of E-Learning Materials

In CANDLE, learning materials are described by metadata and represented in XML (eXtensible Markup Language), following the current paradigm of semantic interoperability among networked information resources. Metadata is simply descriptive information about resources like: title, authors, conditions of use, keywords, technical requirements, etc. Its fundamental role is to enhance the process of information retrieval, by providing rich and machine "understandable" representations.

[1] Collaborative And Network Distributed Learning Environment: http://www.candle.eu.org/

The project is developing its own educational metadata scheme, which specializes IEEE's LOM (Learning Object Metadata) specification [1]. The main purposes of using metadata in CANDLE's context are: -

(a) to aid in the construction of new courseware, re-using already existent materials;
(b) and to enhance the navigation and exploration of the learner.

Among other metadata fields, a category to classify the course materials, according to a pre-defined taxonomy of the telematics domain knowledge has been defined. So, whenever new material is created, its author is asked to select a set of key words from that taxonomy. Additionally, the author is asked to assign a weight to each of the selected key words to quantify their significance to the material being tagged. With this approach, metadata can be used to discover relevant material based on its *knowledge space* location.

The abstract *knowledge space* can be built in several ways. A sophisticated approach to do so is to develop an ontology of the domain, which defines a set of concepts (equivalent to the taxonomic key words) and set of relations between those concepts. So, once learning material is tagged with specific concepts, the relationships defined in the ontology can be used to locate related courseware. This facilitates the search of material both in the authoring and learning scenarios. In the former case, an author might start from one point in the ontology and follow the appropriate links to locate material with the relationships required for his course. More important for the learner the ontology can be used for navigation and possibly automated location of content – by following the relationship links relevant content can be located.

The key question with this approach is which ontology to use. Two problems can be foreseen. On the one hand different experts in a given field are likely to disagree on the correct ontology. On the other hand, in fields like engineering or telematics, an ontology can change through time as the fields develop. Thus, several ontological frameworks need to be available or a dynamic ontology creation process needs to exist.
This leads to an approach supporting a process of discovering the underlying knowledge structure and generating the relationships between learning materials that are part of the courseware database. We are exploring the use of fuzzy clustering to build such dynamic knowledge representation.

4. Fuzzy Clustering for Knowledge Representation

The common goal of clustering methods is to group data elements according to some similarity measure so that related elements are placed in the same cluster. This makes possible the discovery of unobvious relations and structures in data sets.
Information retrieval is one of the many application areas that make use of cluster analysis. Document clustering has long been applied to enhance the process of search and retrieval based on the so-called *cluster hypothesis* [2]. Document clustering has also been used as a post-retrieval tool for browsing large document collections [3].
Our objective is to discover the knowledge-relations that may exist between learning material, based on their metadata descriptions. This can be seen as a document clustering problem. A suitable algorithm may be applied for organizing the XML documents and for discovering hidden relations between them.

Agglomerative hierarchical clustering algorithms are perhaps the most popular for document clustering [4]. These methods have the advantage of providing a hierarchical organization of the document collection, but they are slower than partitional methods, like the K-Means algorithm. For this reason the latter is also very popular. Both methods generate hard clusters, in the sense that each document is assigned to one and only one cluster. Considering our objectives, a method capable of generating fuzzy clusters would be the most suitable. The arguments that support this statement are the following:

(a) The first one concerns the representation of the *knowledge space*. We pointed out previously that a major problem of using an ontology was to be able to define the correct representation of the domain knowledge. Since knowledge is such an abstract thing every attempt to represent it will be to some extent uncertain. Thus, fuzzy document relations are more likely to represent the "true" knowledge structure than crisp ones.

(b) Another reason has to do with the way learning materials are classified. The selection of weighted key words represents the author's best attempt to define the subjects associated each material. But this tagging process is intrinsically imprecise, firstly because authors may differ in their exact

understanding of the key words (which may be occasionally ambiguous) and secondly, the assigned weights express a subjective opinion.

The theory of fuzzy sets provides the mathematical means to deal with uncertainty [5]. Fuzzy clustering brings together the ability to find unobvious relations and structures in data sets with the ability to cope with uncertainty. The following sections present some background and report on the document clustering experiments in which fuzzy clustering techniques have been applied.

5. Background for the Clustering Experiments

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 [6] each document is represented by a set of indexing terms in the form of a *k*-dimensional vector:

$$x_i = [w_{i1}\ w_{i2} \ldots w_{ik}] \tag{5.1}$$

where k is the total number of terms and w_{ij} represents the weight (or significance) of term j in document x_i. The weights w_{ij} can be obtained in various ways. Usually an automatic indexing procedure finds the indexing terms associated with each text document and the weights are then computed as a function of the term frequencies. A term weighting system that has proved to perform well considers the term frequency, the inverse document frequency (that is, the term specificity within the document collection) and a factor of document length normalization [7]:

$$w_{ij} = \frac{f_{ij} \cdot \log(N / n_j)}{\sqrt{\sum_{t=1}^{k} (f_{it} \cdot \log(N / n_t))^2}} \tag{5.2}$$

where f_{ij} is the frequency of term j in document i, n_j is the number of documents that contain term j and N is the total number of documents.
A similar vector representation can be obtained for CANDLE's learning materials from their metadata descriptions. In this case, the indexing terms (key words from the taxonomy) and the associated weights are assigned manually by authors.

The weighted document vectors are suitable to be processed by a clustering algorithm. Regardless of the algorithm, documents will be grouped according to their similarities. Hence, it is necessary to choose an appropriate similarity measure for the document space. A familiar function that is used in the VSM to compare document vectors with query vectors is the based on inner product of the two vectors [7]:

$$S(x_\alpha, x_\beta) = \sum_{j=1}^{k} w_{\alpha j} \cdot w_{\beta j} = x_\alpha \cdot x_\beta{}^T \tag{5.3}$$

Since x_α and x_β are normalized weighted vectors the similarity function exhibits the following properties:

$$0 \le S(x_\alpha, x_\beta) \le 1, \forall_{\alpha,\beta} \tag{5.4}$$

$$S(x_\alpha, x_\alpha) = 1, \forall_\alpha \tag{5.5}$$

Some clustering algorithms group data elements based on dissimilarities or distances. A dissimilarity function can be obtained from the similarity measure defined in (5.3) by an appropriate transformation:

$$D(x_\alpha, x_\beta) = 1 - S(x_\alpha, x_\beta) = 1 - \sum_{j=1}^{k} w_{\alpha j} \cdot w_{\beta j} \tag{5.6}$$

Both the document vector representation and the similarity function introduced above were applied in our document clustering experiments. Next, we present the clustering techniques that were used.

5.1 The Fuzzy C-Means Algorithm

In section 4 we presented the arguments that supported the use of a clustering algorithm capable of generating a fuzzy output. Thus, we had to select one such algorithm.
The Fuzzy C-Means (FCM) [8] 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 it is required in our case. We decided to use this algorithm in our experiments, for its simplicity and for being the fuzzy extension of a technique that is common for document clustering. Furthermore, a study presented in [9] indicates that the FCM can perform at least as well as the traditionally used agglomerative hierarchical clustering method.

The algorithm is summarized as follows. Given a data set with N elements each represented by k-dimensional feature vector, the FCM takes as input a ($N \times k$) matrix $X=[x_i]$. It requires the prior definition of the final number of clusters c ($1<c<N$), the choice of the fuzzification parameter m ($m>1$) and the selection of a distance function $\|\cdot\|$, the most common being the Euclidean norm:

$$d_{i\alpha}{}^2 = \|x_i - v_\alpha\|^2 = \sum_{j=1}^{k} (x_{ij} - v_{\alpha j})^2 \tag{5.7}$$

The algorithm runs iteratively to obtain the cluster centers – $V=[v_\alpha]$: ($c \times k$) – 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 following objective function:

$$J_m(U,V)=\sum_{i=1}^{N}\sum_{\alpha=1}^{c}{u_{\alpha i}}^m {d_{i\alpha}}^2=\sum_{i=1}^{N}\sum_{\alpha=1}^{c}{u_{\alpha i}}^m\left\|x_i-v_\alpha\right\|^2 \quad (5.8)$$

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

1. $u_{\alpha i}\in[0,1],\ \forall_{\alpha\in\{1,\ldots,c\}}\ \forall_{i\in\{1,\ldots,N\}}$ (5.9)

2. $\sum_{\alpha=1}^{c}u_{\alpha i}=1,\ \forall_{i\in\{1,\ldots,N\}}$ (5.10)

3. $0<\sum_{i=1}^{N}u_{\alpha i}<N,\ \forall_{\alpha\in\{1,\ldots,c\}}$ (5.11)

At each iteration, the cluster centers and the grades of membership are updated according to (5.12) and (5.13) respectively:

$$v_\alpha=\frac{\sum_{i=1}^{N}{u_{\alpha i}}^m\cdot x_i}{\sum_{i=1}^{N}{u_{\alpha i}}^m} \quad (5.12)$$

$$u_{\alpha i}=\frac{1}{\sum_{\beta=1}^{c}\left(\frac{{d_{i\alpha}}^2}{{d_{i\beta}}^2}\right)^{1/(m-1)}}=\frac{1}{\sum_{\beta=1}^{c}\left(\frac{\left\|x_i-v_\alpha\right\|}{\left\|x_i-v_\beta\right\|}\right)^{2/(m-1)}} \quad (5.13)$$

The algorithm ends when a termination criterion is met or the maximum number of iterations is achieved.

5.2 The 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. This statement can be supported by the following example. Let us suppose that we have two documents x_A and x_B that are indexed with a set k of terms T. Let us also assume that most of the terms in T, say k', appear neither in x_A nor in x_B. Let us also assume that x_A and x_B have no terms in common. Since the two document vectors agree in k' dimensions in which they both have zero term weights, their Euclidean distance will be relatively small, when in fact x_A and x_B are totally dissimilar.

So, the problem with the Euclidean norm is that the non-occurrence of the same terms in both documents is treated in the same way as the co-occurrence of terms.

A suitable dissimilarity function for document vectors was introduced in (5.6). For the previous example this function results in the maximum value possible, that is $D(x_A,x_B)=1$, indicating total dissimilarity.

Thus, we decided to apply the dissimilarity measure as the metric for clustering documents, using the Fuzzy C-Means approach. The modified objective function is similar to (5.8), but now the norm $\|\cdot\|^2$ is replaced by the function defined in (5.6):

$$J_m(U,V)=\sum_{i=1}^{N}\sum_{\alpha=1}^{c}{u_{\alpha i}}^m D_{i\alpha}=\sum_{i=1}^{N}\sum_{\alpha=1}^{c}{u_{\alpha i}}^m\left(1-\sum_{j=1}^{k}x_{ij}\cdot v_{\alpha j}\right) \quad (5.14)$$

As the expression used to update of the clusters centers (5.12) was obtained considering the Euclidean distance we had to derive a new expression to work with the new metric. In order to use the dissimilarity measure, such that property (5.5) holds, the cluster centers need to be normalized. Therefore, we had to introduce the following constraint:

$$S(v_\alpha,v_\alpha)=\sum_{j=1}^{k}v_{\alpha j}\cdot v_{\alpha j}=\sum_{j=1}^{k}{v_{\alpha j}}^2=1,\forall\alpha \quad (5.15)$$

It can be proved that minimizing (5.14) with respect to $u_{\alpha i}$ leads to a similar result as in (5.13), but now ${d_{i\alpha}}^2$ and ${d_{i\beta}}^2$ are replaced by $D_{i\alpha}$ and $D_{i\beta}$,. The expression for $u_{\alpha i}$ is:

$$u_{\alpha i}=\frac{1}{\sum_{\beta=1}^{c}\left(\frac{D_{i\alpha}}{D_{i\beta}}\right)^{1/(m-1)}}=\frac{1}{\sum_{\beta=1}^{c}\left(\frac{1-\sum_{j=1}^{k}x_{ij}\cdot v_{\alpha j}}{1-\sum_{j=1}^{k}x_{ij}\cdot v_{\beta j}}\right)^{1/(m-1)}} \quad (5.16)$$

To minimize (5.14) with respect to v_α we applied the method of the Lagrange multipliers to introduce the constraint (5.15), obtaining:

$$L=\sum_{i=1}^{N}\sum_{\alpha=1}^{c}{u_{\alpha i}}^m\left(1-\sum_{j=1}^{k}x_{ij}\cdot v_{\alpha j}\right)+\sum_{\alpha=1}^{c}\lambda_\alpha\left(\sum_{j=1}^{k}{v_{\alpha j}}^2-1\right) \quad (5.17)$$

Then,

$$\frac{\partial L}{\partial v_\alpha}=-\sum_{i=1}^{N}{u_{\alpha i}}^m x_i+2\lambda_\alpha v_\alpha=0\Leftrightarrow v_\alpha=\frac{1}{2\lambda_\alpha}\cdot\sum_{i=1}^{N}{u_{\alpha i}}^m x_i \quad (5.18)$$

By applying the constraint (5.15) we get

$$\sum_{j=1}^{k}{v_{\alpha j}}^2=\left(\frac{1}{2\lambda_\alpha}\right)^2\cdot\sum_{j=1}^{k}\left(\sum_{i=1}^{N}{u_{\alpha i}}^m x_{ij}\right)^2=1\Leftrightarrow$$

$$\Leftrightarrow\frac{1}{2\lambda_\alpha}=\sqrt{\frac{1}{\sum_{j=1}^{k}\left(\sum_{i=1}^{N}{u_{\alpha i}}^m x_{ij}\right)^2}} \quad (5.19)$$

Replacing $\frac{1}{2\lambda_\alpha}$ in (5.18) we obtain

$$v_\alpha = \sum_{i=1}^{N} {u_{\alpha i}}^m x_i \cdot \sqrt{\frac{1}{\sum_{j=1}^{k}\left(\sum_{i=1}^{N} {u_{\alpha i}}^m x_{ij}\right)^2}} \qquad (5.20)$$

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

5.3 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 4, 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*), which is defined as [8]:

$$PE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{\alpha=1}^{c} u_{\alpha i}\log_a(u_{\alpha i}) \qquad (5.21)$$

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 ($u_{\alpha i} = 1/c$).

6. Experiments with Fuzzy Clustering

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.

6.1 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 [10]. Document vectors as in (5.1) 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 [10]. 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 [11].

6.2 Experimental Results

In our first trials the objective was to analyze if the FCM algorithm would be able to generate a good partition of the document collection. We fixed the number of clusters in 6 (as our benchmark indicated) and we ran the algorithm applying both the Euclidean distance (FCM-ED) and the dissimilarity function (FCM-DF). For each case we tried several values of the fuzzification parameter – $m \in [1.1,2.5]$. 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 (5.2). For the FCM-ED trials we also generated a matrix of term frequency counts ($X_2=[f_{ij}]$).

To be able to compare the results with the reference clustering we used the maximum membership criterion to generate hard clusters from the fuzzy ones. Initially we set m=1.1 so that the results would be close to the hard case. When X_1 was used as input, both FCM-ED and FCM-DF performed quite well generating clusters with a high degree of match with the benchmark. We found out that the FCM-ED performed poorly when X_2 was used even for such a low value of *m*, ~86% of the documents ending up in the same cluster. We also noticed that in this case for increasing values of *m* the execution times increased exponentially. When X_1 was used as input, the computation times were fairly stable for increasing values of *m*, both with the FCM-ED and the FCM-DF. This result is shown in Figure 1. From the plot we see that the FCM-DF converges faster than FCM-ED for $m \leq 1.3$ and slower for $m \geq 1.4$. Although this suggests that for increasing fuzziness the FCM-ED has lower execution times, these times just decrease because the maximum fuzziness has been achieved. In Figure 2 evidence of this is presented. We can observe that for $m \geq 1.4$ the partition entropy is maximal.

An important remark is that the FCM-DF successfully obtains fuzzier partitions. For higher values of *m* the matching between benchmark and hardened clusters is lower, but the partitions generated are still fairly good.

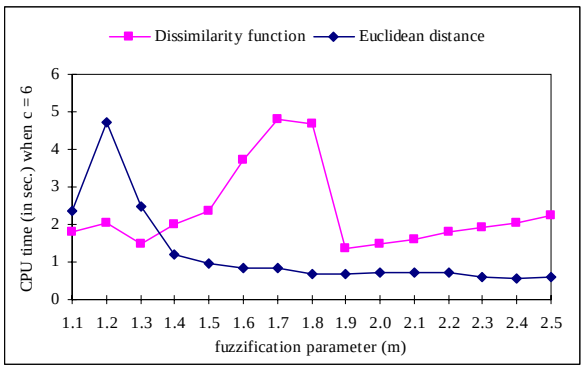

Figure 1. Comparison of the computation times for increasing values for *m*, with *c* set to 6 clusters, using X_1 as input data

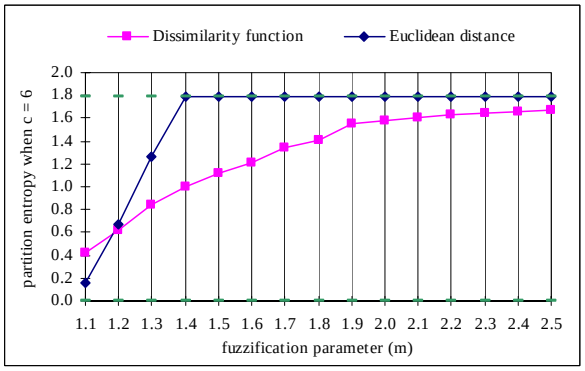

Figure 2. Comparison of the partition entropy for increasing values for *m*, with *c* set to 6 clusters, using X_1 as input data

The decrease verified on the execution times when *m* goes from 1.8 to 1.9 is due to the fact that the partition entropy has increased. This is not very evident from the plot but in fact, when *m*=1.8 around 28% of the documents still have maximum membership ≥0.5 and only 7% of them have maximum membership close to $1/c \cong 0.17$. But when m=1.9, the first statistic decreases to 18% and the second one increases to 24%.

7. Conclusions and Future Work

In this paper, we presented an approach to dynamically represent knowledge domains through the discovery of fuzzy relationships between E-Leaning materials. We proposed 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 higher degrees of fuzziness were acceptable without compromising the quality of the clusters. This is an important result that answers our requirements regarding knowledge-based organization of CANDLE's learning materials.

In the near future, our research will address issues regarding the incremental update of the fuzzy clusters to deal with new document arrivals in the database. A hierarchical organization of the learning materials based on a nested refinement of the fuzzy partitions will also be investigated.

8. Acknowledgments

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

9. References

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

[2] C. J. van Rijsbergen. *Information Retrieval*. Second Edition, Butterworth, London, 1979.

[3] D. R. Cutting, D.R. Karger, J.O. Pedersen, J. W. Tukey. "Scatter/Gather: a cluster-based approach to browsing large document collections," SIGIR'92, 1992, pp. 318-329.

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

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

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

[7] 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.

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

[9] D. H. Kraft, J. Chen, A. Mikulcic. "Combining Fuzzy Clustering and Fuzzy Inference in Information Retrieval," FUZZ IEEE 2000, Vol. 1, 2000, pp. 375-380.

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

[11] 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, 2000, London.