Evaluating Fuzzy Clustering for Relevance-based Information Access

M. E. S. Mendes and L. Sacks
Dept. of E&EE, University College London
Torrington Place, London, WC1E 7JE, UK
{mmendes, lsacks}@ee.ucl.ac.uk

***Abstract* – This paper analyzes the suitability of fuzzy clustering methods for the discovery of relevant document relationships, motivated by the need for enhanced relevance-based navigation of Web-accessible resources. The performance evaluation of a modified *Fuzzy c-Means* algorithm is carried out, and a comparison with a traditional hard clustering technique is presented. Clustering *precision* and *recall* are defined and applied as quantitative evaluation measures of the clustering results. The experiments with various test document sets have shown that in most cases fuzzy clustering performs better than the hard *k-Means* algorithm and that the fuzzy membership values can be used to determine document relevance and to control the amount of information retrieved to the user.**

I. Introduction

The goal of every clustering algorithm is to group data elements according to some (dis)similarity measure so that unobvious relations and structures in the data can be revealed. Document clustering techniques have been widely applied in the field of Information Retrieval (IR) for improving search and retrieval efficiency. The use of clustering in this area is supported by the cluster hypothesis [1] which assumes that documents relevant to a given query tend to be more similar to each other than to irrelevant documents and hence are likely to be clustered together. Clustering has also been used as a tool for browsing large document collections [2] and as a post-retrieval tool for organizing Web search results into meaningful groups [3].

Our motivation for using document clustering techniques is to enable relevance-based access to information resources, with particular application to network-based teaching and learning systems - *e*-Learning. In such systems large online repositories of learning material may be accessed by students, but it is necessary to narrow down the available resources to a particular individual based on the learning context, *i.e.* to take into account the student's background knowledge, learning objectives and pedagogical approaches. This may range from relatively rigid training objectives through to exploratory or research oriented interactions. In the two last cases, tools are required to determine which documents are the most relevant for a given student who wants to learn a particular subject.

The calculation of document relevance requires some knowledge about the content relationships, hence there needs to be a way to classify and organize information in terms of knowledge domains. An emerging approach is to develop an ontology of the given domain that defines a set of concepts and relations between those concepts, which are then used to manually classify documents. The rich semantic information captured by the ontology facilitates the search and navigation of content. The ontology approach was proposed for the Semantic Web [4], but a key question that arises 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, fields evolve and the true ontology quickly changes through time as the fields develop. Consequently, the deployment and maintenance efforts are costly. Instead of the static ontology model, we propose a process of dynamic ontology discovery that applies fuzzy clustering to identify document relationships.

The subsequent sections of this paper are organized as follows: In section II, the argument for using fuzzy clustering techniques instead of traditional hard clustering methods is supported and some considerations regarding the choice of a distance function for document collections are presented. The last part of this section contains a modified *Fuzzy c-Means* clustering algorithm that replaces the squared Euclidean norm by a dissimilarity function common to IR systems. In section III, the performance evaluation measures that have been used in our document clustering experiments are introduced. In section IV, the experimental work is described and the results are presented and analyzed. Finally, section V contains the conclusions.

II. Document Clustering

A. Hard vs. fuzzy clustering

Agglomerative hierarchical clustering (AHC) algorithms are perhaps the most popular for document clustering [5]. Such methods have the advantage of providing a hierarchical organization of the document collection but their time complexity is problematic when compared to partitional methods such as the *k-Means* algorithm [6] (also often used for document clustering).

Both AHC and *k-Means* generate hard clusters, meaning that each document is assigned to a single cluster. But, given that our goal is to discover the best representation for the true ontology of a given domain, we explore fuzzy clustering algorithms instead. In general, the concepts that characterize each knowledge domain are somehow associated with each other, but many times those concepts are also related to concepts of different domains. Consequently, documents may contain information that is relevant to different domains to

This work was supported by the Portuguese Foundation for Science and Technology (Fundação para a Ciência e a Tecnologia) through the doctoral scholarship programme (grant ref. PRAXIS XXI/ BD/21768/99).

some degree. With fuzzy clustering, documents may be attributed to several clusters simultaneously and so, useful relationships between domains may be uncovered, which would otherwise be neglected by hard clustering methods. Moreover, fuzzy clustering methods like the *Fuzzy c-Means* (FCM) algorithm [7] generate fuzzy weights that represent the degree of membership of each data element/document in each cluster. Such weights may be used to obtain fuzzy relations between documents and to determine document relevance.

Although fuzzy clustering has not been widely explored for document clustering, some recent research in this area has been carried out [8][9][10][11][12][13]. In our study, we have decided to use the FCM algorithm due to its simplicity and for being the soft version of the *k-Means* algorithm that has long been used for document clustering.

B. Selection of a distance function

The choice of a particular distance function to be used in clustering algorithms should reflect the nature of the data set. Documents are usually represented as term vectors according to the Vector Space model of IR [14] and those vectors tend to be high-dimensional and very sparse. The Euclidean distance, which is commonly applied in the FCM algorithm, is not the most suitable metric for measuring the proximity between documents. The problem with this norm is that the non-occurrence of the same terms in both documents is handled in the similar way as the co-occurrence of terms. Measures like the cosine similarity [14] from the field of IR, are better suited to determine the proximity of documents. The cosine measure, denoted here as $S_{\alpha\beta}$ (1), is simply the inner product of k-dimensional vectors (x_α and x_β) after normalization to unit length (*i.e.* $\|x_\alpha\|=\|x_\beta\|=1$). The higher the cosine value the higher the similarity between the documents.

$$S(x_\alpha, x_\beta) = \left\langle x_\alpha, x_\beta \right\rangle = \sum_{j=1}^{k} x_{\alpha j} \cdot x_{\beta j} \tag{1}$$

The similarity measure exhibits properties (2) and (3):

$$0 \le S(x_\alpha, x_\beta) \le 1, \forall_{\alpha,\beta} \quad (2) \qquad S(x_\alpha, x_\alpha) = 1, \forall_\alpha \quad (3)$$

A simple transformation to (1) can be performed to obtain the dissimilarity function in (4), with properties (5) and (6).

$$D(x_\alpha, x_\beta) = 1 - S(x_\alpha, x_\beta) = 1 - \sum_{j=1}^{k} x_{\alpha j} \cdot x_{\beta j} \tag{4}$$

$$0 \le D(x_\alpha, x_\beta) \le 1, \forall_{\alpha,\beta} \quad (5) \qquad D(x_\alpha, x_\alpha) = 0, \forall_\alpha \quad (6)$$

Since this function is known to work better for document vectors than the Euclidean distance, we have selected it. The chosen fuzzy clustering algorithm (FCM) had to be modified to use the dissimilarity function above. Such modification is presented in the next sub-section.

C. Hyperspherical Fuzzy c-Means algorithm

We have recently proposed [9] applying the dissimilarity function (4) instead of the Euclidean distance for clustering normalized document vectors using the FCM approach. Similar use of the dissimilarity function in fuzzy clustering was also explored in [10] and [15]. Our previous experiments proved that with the dissimilarity function significantly better results were achieved.

A modification of the original objective function was required and therefore a new expression for updating the cluster centers had to be defined. The modified algorithm has been labeled *Hyperspherical Fuzzy c-Means* (H-FCM), as both data vectors and cluster centers lie in a k-dimensional hypersphere of unit radius.

The modified objective function (7) is similar to the original one, the difference being the replacement of the squared norm by the function defined in (4):

$$J_m(U,V) = \sum_{i=1}^{N} \sum_{\alpha=1}^{c} u_{\alpha i}{}^m D_{i\alpha} = \sum_{i=1}^{N} \sum_{\alpha=1}^{c} u_{\alpha i}{}^m (1 - \sum_{j=1}^{k} x_{ij} \cdot v_{\alpha j}) \,. \tag{7}$$

The constraints regarding the membership values $u_{\alpha i}$ are the same as those in the original FCM and the update expression for the membership values (8) is also similar to the original one since the calculation of $D_{i\alpha}$ does not depend explicitly of $u_{\alpha i}$:

$$u_{\alpha i} = \sum_{\beta=1}^{c} \left(\frac{D_{i\alpha}}{D_{i\beta}} \right)^{-\frac{1}{(m-1)}} = \sum_{\beta=1}^{c} \left(\frac{1 - \sum_{j=1}^{k} x_{ij} \cdot v_{\alpha j}}{1 - \sum_{j=1}^{k} x_{ij} \cdot v_{\beta j}} \right)^{-\frac{1}{(m-1)}} \,. \tag{8}$$

The constraint for the cluster prototype vectors v_α in (9) was introduced so that properties (5) and (6) would hold for every $D_{i\alpha}$:

$$S(v_\alpha, v_\alpha) = \sum_{j=1}^{k} v_{\alpha j} \cdot v_{\alpha j} = \sum_{j=1}^{k} v_{\alpha j}{}^2 = 1, \forall_\alpha \,. \tag{9}$$

This constraint forces the cluster centers to be normalized to unit length. The new update expression for the centers was derived by minimizing (7) with respect to v_α ($u_{\alpha i}$ fixed) subject to constraint (9), using the method of the Lagrange multipliers. The Lagrangian function is defined as:

$$\begin{aligned} L(v_\alpha, \lambda_\alpha) &= J_m(U, v_\alpha) + \lambda_\alpha \cdot [S(v_\alpha, v_\alpha) - 1] \\ &= \sum_{i=1}^{N} u_{\alpha i}{}^m (1 - \sum_{j=1}^{k} x_{ij} \cdot v_{\alpha j}) + \lambda_\alpha (\sum_{j=1}^{k} v_{\alpha j}{}^2 - 1) \end{aligned} \tag{10}$$

where λ_α is the Lagrange multiplier. This minimization problem is converted into an unconstrained problem taking the derivative of the Lagrangian function,

$$\frac{\partial L(v_\alpha, \lambda_\alpha)}{\partial v_\alpha} = \frac{\partial J_m(U, v_\alpha)}{\partial v_\alpha} + \lambda_\alpha \cdot \frac{\partial [S(v_\alpha, v_\alpha) - 1]}{\partial v_\alpha} = 0 \tag{11}$$

which is equivalent to,

$$-\sum_{i=1}^{N} u_{\alpha i}{}^{m} x_i + 2\lambda_\alpha v_\alpha = 0 \Leftrightarrow v_\alpha = \frac{1}{2\lambda_\alpha} \cdot \sum_{i=1}^{N} u_{\alpha i}{}^{m} x_i \ . \tag{12}$$

Applying constraint (9) follows,

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

$$\Leftrightarrow \frac{1}{2\lambda_\alpha} = \left[\sum_{j=1}^{k} \left(\sum_{i=1}^{N} u_{\alpha i}{}^{m} x_{ij}\right)^2\right]^{-1/2} \tag{13}$$

and replacing $\frac{1}{2\lambda_\alpha}$ in (12) leads to,

$$v_\alpha = \sum_{i=1}^{N} u_{\alpha i}{}^{m} x_i \cdot \left[\sum_{j=1}^{k} \left(\sum_{i=1}^{N} u_{\alpha i}{}^{m} x_{ij}\right)^2\right]^{-1/2} \tag{14}$$

Like the original algorithm, H-FCM runs iteratively until a local minimum of the objective function is found or the maximum number of iterations is reached.

III. Performance Evaluation

The validity of fuzzy clustering algorithms is generally evaluated using internal performance measures, *i.e.* measures that are algorithm dependent and do not contain any external or objective knowledge about the actual structure of the data set. This is the case of various validity indexes for the FCM algorithm such as the Partition Entropy [7], the Xie-Beni index [16] or the Fukuyama-Sugeno index [17]. When there is prior knowledge on how clusters should be formed external performance measures (algorithm independent) can be used to compare the clustering results with the benchmark.

Two popular measures that are typically used to evaluate the performance of IR systems are *precision* and *recall* [1][14]. In such systems *precision* represents the fraction of relevant documents out of those retrieved in response to a particular query and *recall* represents the fraction of retrieved documents out of the relevant ones. Similar measures have been applied for the evaluation of classification systems [18], whose purpose is to classify data elements given a known set of classes. In this case, *precision* represents the fraction of elements assigned to a pre-defined class that indeed belong to that class and *recall* represents the fraction of elements that belong to a pre-defined class that were effectively assigned to that class. Likewise, *precision* and *recall* can be used as external performance measures for evaluating clustering algorithms (that are in fact unsupervised classification systems) in cases where a clustering benchmark exists.

Given a discovered cluster γ and the associated reference cluster Γ, *precision* ($P_{\gamma\Gamma}$) and *recall* ($R_{\gamma\Gamma}$) are defined as follows:

$$P_{\gamma\Gamma} = \frac{n_{\gamma\Gamma}}{N_\gamma} \ , \tag{15}$$

$$R_{\gamma\Gamma} = \frac{n_{\gamma\Gamma}}{N_\Gamma} \ , \tag{16}$$

where $n_{\gamma\Gamma}$ is the number of documents from reference cluster Γ assigned to cluster γ, N_γ is the total number of documents in cluster γ and N_Γ is the total number of documents in reference cluster Γ. These two performance measures can be combined into a single measure, the *F-measure* [1][19], that is defined as:

$$F^{\xi}{}_{\gamma\Gamma} = \frac{(\xi^2+1) \cdot P_{\gamma\Gamma} \cdot R_{\gamma\Gamma}}{\xi^2 \cdot P_{\gamma\Gamma} + R_{\gamma\Gamma}} \ , \tag{17}$$

where ξ is a parameter that controls the relative weight of *precision* and *recall* (ξ=1 is used for equal contribution). To obtain overall performance measures, a weighted average of the individual $P_{\gamma\Gamma}$ and $R_{\gamma\Gamma}$ is applied:

$$P = \frac{\sum_{\Gamma=1}^{c} N_\Gamma P_{\gamma\Gamma}}{\sum_{\Gamma=1}^{c} N_\Gamma} \ , \tag{18}$$

$$R = \frac{\sum_{\Gamma=1}^{c} N_\Gamma R_{\gamma\Gamma}}{\sum_{\Gamma=1}^{c} N_\Gamma} \ . \tag{19}$$

The measures that have just been described consider hard clusters. In the fuzzy clustering case, documents may have membership in multiple clusters and it is even possible that all documents belong to some degree to all clusters. In such case *precision* would be consequently low. Hence, either a soft version of the measures is defined – *fuzzy precision* and *fuzzy recall* – or the fuzzy clusters are made crisp before calculating the measures, using for instance the maximum membership criterion. In the work presented in this paper, we have hardened the clusters for various membership thresholds (α-cuts) and calculated $P_{\gamma\Gamma}$ and $R_{\gamma\Gamma}$ for each case.

IV. Experimental Trials

A. Description of the Data Sets

Three different collections were selected for the document clustering experiments: the *Reuters-21578* text categorization collection[2], a subset of the *Open Directory Project* (ODP) metadata[3] and scientific abstracts obtained from the INSPEC database[4]:

- The *Reuters*-21578 text collection consists of newswire articles classified into 135 topic categories. We have selected articles belonging to at least one topic, using the "ModApte" split (*i.e.* LEWISSPLIT = "TEST" and TOPICS = "YES"). Two subsets were generated for the most frequent topics in the collection: *reuters1*, a subset containing articles classified with a single topic - "trade", "acq" or "earn" - and *reuters2*, a subset containing

[2] Reuters-21578 test collection: http://www.daviddlewis.com/resources/testcollections/

[3] Open Directory Project (ODP) : http://dmoz.org/

[4] INSPEC database: http://www.iee.org/publish/inspec/

articles classified with one or more topics - “money-fx”, “ship”, “interest”, “trade” and “crude”.

- The ODP is a human-edited directory of the World Wide Web, where Web sites are categorized into a topic hierarchy and represented by metadata in the RDF format [20]. A subset of the directory was selected, the *Kids and Teens* topic hierarchy, and we have created the *odp* test collection with the short metadata descriptions of Web sites related to the following topics: “drugs”, “health” and “sports”.
- The INSPEC database is a scientific database of abstracts in the fields of physics, electronics and electrical engineering, computers and control, and information technology. We have generated two test sets *inspec1* and *inspec2* by downloading all the abstracts published since 2000 and classified with the following topics: “back-propagation”, “fuzzy control” and “pattern classification” (*inspec1*) and “broadband network”, “multimedia communication” and “queueing theory” (*inspec2*).

The size of each document collection and the distribution of documents per topic are shown in Table I.

TABLE I

DATA SETS DESCRIPTION

Collection ($N \times k$)	Topics	No. docs in this topic	No. docs only in this topic
reuters1 (908×10582)	trade	410	410
	acq	247	247
	earn	251	251
reuters2 (1374×11778)	money-fx	343	253
	ship	440	190
	interest	488	206
	trade	194	108
	crude	299	251
odp (404×551)	drugs	48	44
	health	103	95
	sport	262	256
inspec1 (7971×11803)	backpropagation	2271	2174
	fuzzy control	3899	3800
	pattern classification	1920	1879
inspec2 (9082×13782)	broadband network	2773	2296
	multimedia communication	3748	3234
	queueing theory	3185	2951

B. Document representation

Each document was automatically indexed for keyword frequency extraction. Stemming was performed (*i.e.* word affixes such as ‘ing’, ‘ion’, ‘s’, were removed) [21] and stop words were discarded (*i.e.* insignificant words like ‘a’, ‘and’, ‘where’, ‘or’) [22]. Documents were represented as *tf* (term frequency) vectors according to the Vector Space model of IR [14]. The vectors were then organized as rows of a ($N \times k$) matrix, where N is the collection size and k is the total number of indexing terms (Table I contains the specific values of N and k for each collection).

C. Experiments and results

The main goals of the document clustering experiments were to investigate the suitability of fuzzy clustering for discovering good document relationships by assessing the quality of the obtained clusters and to compare this approach with a traditional hard clustering technique.

For each test collection we set the number of clusters c equal to the number of topics in Table I. We run both the *k-Means* and the H-FCM algorithm for each collection. From the results a confusion matrix was obtained and from the analysis of this matrix we were able to identify a correspondence between found clusters and reference clusters. In the *k-Means* case, *precision* (P) and *recall* (R) were calculated for each cluster and averaged to obtain an overall value (see section III). The same procedure was followed in the H-FCM case but the individual measures were calculated for various α-cuts of the partition matrix, *i.e.* documents with membership value in a given cluster above a fixed threshold α were attributed to that cluster to then calculate P and R. The graphs in Figs. 1 to 5 contain the results of the *k-Means* and H-FCM for each collection. The H-FCM data in these plots refers to the case when m was set to 1.10. Such a low value of m was used to approximate the fuzzy clusters to the crisp case (since as m tends to 1 the fuzzy partition tends to a hard partition).

It is desirable that both P and R are as high as possible. Ideally they would both be equal to 1, which would mean that every cluster contained all and only the right documents. For different collections the maximum values obtained for P and R varied. An important result is that for the same level of R, the H-FCM achieved higher P than the *k-Means*, with 4 collections (see Figs. 1, 2, 4 and 5). With the *odp* collection that does not happen, but if a α-cut between 0.2 or 0.3 is applied, the same level of R is possible with just a small difference in P of around 0.05 (see Fig. 3).

The H-FCM algorithm was also run for higher values of the fuzzification parameter m. As expected, for the same α-cut it was observed that with increasing m, *recall* was generally higher and *precision* was lower.

A great advantage of the H-FCM is that *precision* and *recall* can be controlled by setting different thresholds for α. It is obvious that lowering the threshold will lead to more documents being attributed simultaneously to more clusters, hence increasing R and decreasing P. The *F-measure* can be used to decide which α-cut leads to the best compromise between P and R, *i.e.* when F is maximized. The H-FCM results in Figs. 1 to 5 point out which α-cut maximizes F.

Another advantage of algorithms that compute a centre for each cluster is that these prototypes are themselves term vectors that can be used for automatic labeling of the cluster contents. To illustrate, Table II contains the top ten terms and

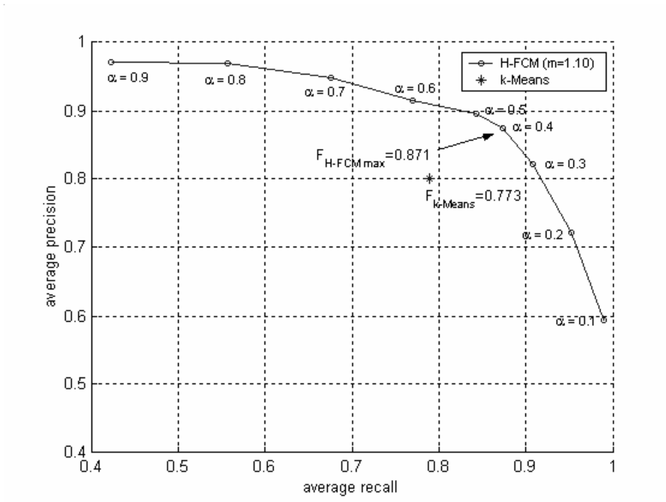

Figure 1 – Average *precision* vs. *recall* obtained for the *reuters1* collection

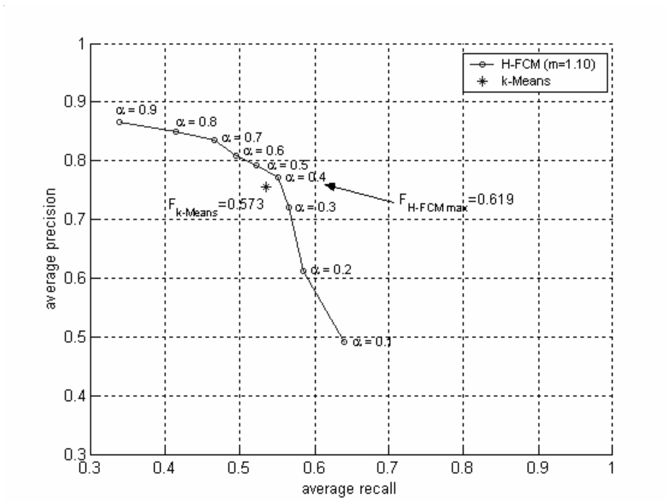

Figure 2 – Average *precision* vs. *recall* obtained for the *reuters2* collection

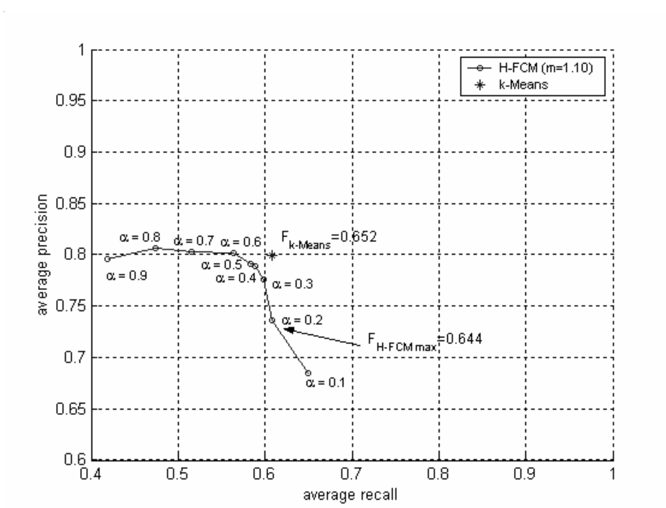

Figure 3 – Average *precision* vs. *recall* obtained for the *odp* collection

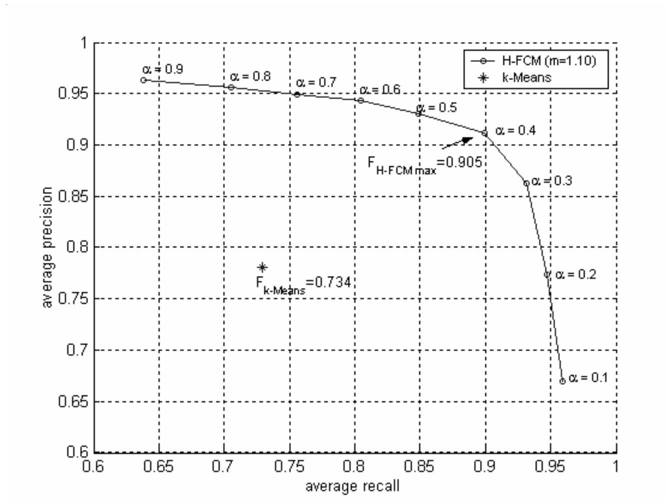

Figure 4 – Average *precision* vs. *recall* obtained for the *inspec1* collection

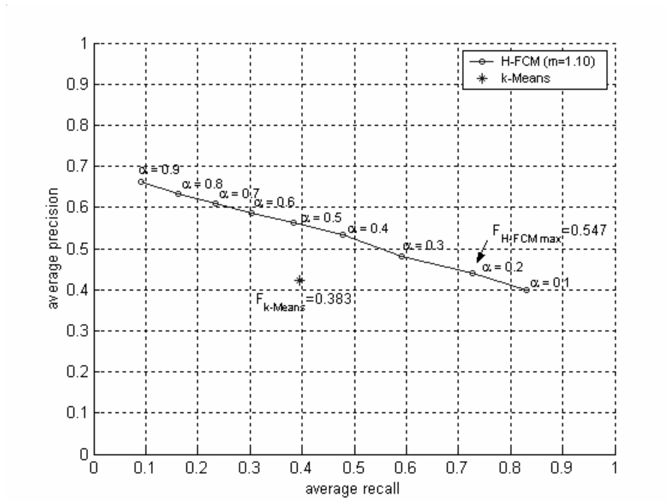

Figure 5 – Average *precision* vs. *recall* obtained for the *inspec2* collection

respective weights of the cluster centers obtained with the H-FCM algorithm for each collection. There is a fairly good correspondence between the terms and the topics in Table I.

V. Conclusions

Fuzzy clustering has been studied for the discovery of document relationships to support relevance-based access and flexible exploration of *e*-Learning content. Considering the requirements of our application, fuzzy methods present some advantages over traditional document clustering techniques that generate crisp partitions. Therefore, several experiments were carried out with different test collections to compare the performance of the *Hyperspherical Fuzzy c-Means* (H-FCM) algorithm to that of the well-known *k-Means*. *Precision* and *recall* were used as objective quantitative measures of the clusters quality. Our study has shown that in most cases the performance of the H-FCM is superior to that of the *k-Means*.

TABLE II

TOP TEN TERMS AND WEIGHTS OF THE CLUSTER CENTERS DISCOVERED BY H-FCM (WITH m=1.10)

Collection	Cluster Centers
reuters1 (3 clusters)	trade (0.642), blah (0.289), japan (0.249), billion (0.181), reuter (0.161), march (0.157), japanese (0.143), year (0.121), dlrs (0.112), countries (0.095)
	dlrs (0.357), march (0.308), reuter (0.305), company (0.292), mln (0.255), pct (0.242), corp (0.222), shares (0.203), stock (0.168), offer (0.142)
	mln (0.472), cts (0.441), net (0.284), march (0.266), reuter (0.259), loss (0.238), dlrs (0.233), shr (0.176), profit (0.141), year (0.141)
reuters2 (5 clusters)	blah (0.914), pct (0.150), rate (0.128), fed (0.122), bank (0.112), trade (0.086), billion (0.081), sets (0.074), repurchase (0.072), customer (0.064)
	mln (0.520), stg (0.476), bank (0.331), market (0.275), money (0.227), reuter (0.153), pct (0.151), march (0.148), today (0.133), england (0.127)
	pct (0.489), rate (0.299), bank (0.296), reuter (0.222), march (0.210), market (0.195), billion (0.174), rates (0.174), fed (0.173), federal (0.133)
	trade (0.589), japan (0.279), reuter (0.185), march (0.183), billion (0.164), japanese (0.140), year (0.131), washington (0.118), countries (0.115), told (0.106)
	oil (0.641), march (0.227), reuter (0.210), dlrs (0.169), crude (0.162), mln (0.162), opec (0.150), prices (0.141), pct (0.108), bpd (0.102)
odp (3 clusters)	teen (0.611), health (0.559), drug (0.266), kid (0.244), inform (0.162), top (0.112), sexual (0.097), life (0.097), educ (0.085), includ (0.084)
	camp (0.786), sport (0.453), summer (0.146), dai (0.124), locat (0.114), art (0.113), ag (0.112), activ (0.100), program (0.096), kid (0.082)
	sport (0.854), kid (0.289), teen (0.175), top (0.124), game (0.122), includ (0.102), inform (0.096), histori (0.084), featur (0.064), olymp (0.063)
inspec1 (3 clusters)	network (0.650), neural (0.540), algorithm (0.156), model (0.149), system (0.139), base (0.137), learn (0.119), method (0.114), train (0.097), backpropag (0.086)
	control (0.723), fuzzi (0.529), system (0.266), base (0.116), model (0.099), logic (0.090), adapt (0.088), design (0.082), method (0.064), nonlinear (0.056)
	cluster (0.744), algorithm (0.276), data (0.243), base (0.185), imag (0.170), fuzzi (0.169), method (0.151), model (0.114), approach (0.081), analysi (0.079)
inspec2 (3 clusters)	network (0.737), servic (0.265), multimedia (0.159), wireless (0.141), broadband (0.126), base (0.117), access (0.109), control (0.106), traffic (0.103), atm (0.102)
	system (0.641), multimedia (0.231), commun (0.185), servic (0.156), wireless (0.154), cdma (0.146), perform (0.144), channel (0.139), base (0.136), mobil (0.131)
	queue (0.332), model (0.252), servic (0.250), traffic (0.232), time (0.212), network (0.212), perform (0.179), control (0.167), system (0.165), base (0.160)

Moreover, H-FCM has the advantage of generating cluster membership values thereby attributing documents to multiple clusters simultaneously. Such a characteristic is particularly important in applications like ours where documents may be relevant to different knowledge domains to some degree. Finally, another important advantage of having a fuzzy partition is that *precision* and *recall* can be tuned by applying different α-cuts for the membership values. The significance of this result is better understood considering a cluster-based search tool, where the user would be able to control the number of documents to be displayed depending on his/her browsing objectives.

REFERENCES

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

[2] D. R. Cutting, D. R.Karger, J. O. Pedersen and J. W. Tukey JW, "Scatter/Gather: a cluster-based approach to browsing large document collections," *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'92, pp. 318-329, Jun. 1992.

[3] O. Zamir and O. Etzioni, "Grouper: a dynamic clustering interface to Web search results," *Computer Networks*, vol. 31, no. 11-16, pp. 1361-1374, May 1999.

[4] T. Berners-Lee, J. Hendler, O. Lassila, "The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities," *Scientific American*, vol. 284, no. 5, pp. 34-43, May 2001.

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

[6] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the 5th Berkeley Symposium on Mathematics, Statistics and Probability*, vol. 1, pp. 281-296, 1967.

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

[8] D. H. Kraft, J. Chen and A. Mikulcic, "Combining fuzzy clustering and fuzzy inference in information retrieval," *Proceedings of the 9th IEEE International Conference on Fuzzy Systems*, FUZZ IEEE 2000, vol. 1, pp. 375-380, May 2000.

[9] M. E. S. Mendes and L. Sacks, "Dynamic knowledge representation for *e*-Learning applications," *Proceedings of the 2001 BISC International Workshop on Fuzzy Logic and the Internet*, FLINT 2001, Memorandum No. UCB/ERL M01/28, pp. 176-181, U. C. Berkeley, Aug. 2001.

[10] S. Miyamoto, "Fuzzy multisets and fuzzy clustering of documents," *Proceedings of the 10th IEEE International Conference on Fuzzy Systems*, FUZZ IEEE 2001, vol. 2, pp. 1191-1194, Dec. 2001.

[11] H. Frigui and O. Nasraoui, "Simultaneous categorization of text documents and identification of cluster-dependent keywords," *Proceedings of the 2002 IEEE International Conference on Fuzzy Systems*, FUZZ IEEE 2002, vol. 2, pp. 1108-1113, May 2002.

[12] R. Kondadadi and R. Kozma, "A modified fuzzy ART for soft document clustering", *Proceedings of the 2002 International Joint Conference on Neural Networks*, IJCNN '02, vol. 3, pp. 2545- 2549, 2002.

[13] R. Krishnapuram, A. Joshi and Liyu Yi, "A fuzzy relative of the k-Medoids algorithm with application to web document and snippet clustering," *Proceedings of the1999 IEEE International Conference on Fuzzy Systems*, FUZZ IEEE 1999, vol. 3, pp. 1281-1286, Aug. 1999.

[14] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. New York: Addison Wesley, ACM Press, 1999.

[15] F. Klawonn, A. Keller, "Fuzzy clustering based on modified distance measures," *Proceedings of the Third International Symposium on Intelligent Data Analysis*, IDA'99, LNCS 1642, pp. 291-301, Aug. 1999.

[16] X. L. Xie and G. A. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841-847, Aug. 1991.

[17] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method," *Procceedings of the 5th Fuzzy Systems Symposium*, pp. 247-250, 1989.

[18] D. D. Lewis, "Evaluating text categorization," *Proceedings of the Speech and Natural Language Workshop*, pp. 312-318, Feb. 1991.

[19] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 94, pp. 3-12, Aug. 1994.

[20] O. Lassila and R.R. Swick, "Resource Description Framework (RDF) - Model and Syntax Specification," *W3C Recommendation*, Feb. 1999. Available at: http://www.w3.org/ TR/REC-rdf-syntax/

[21] G. Salton, *A Theory of Indexing*. Philadelphia: Society for Industrial and Applied Mathematics, 1975.

[22] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130-137, Jul. 1980.