

Proposal of a Document Cluster Representation based on Fuzzy Deformable Prototypes

Francisco P. Romero
Soluziona Software Factory
R+D Center Soluziona – UCLM
Ronda de Toledo, s/n
13071-Ciudad Real, SPAIN
fpromero@soluziona.com

José A. Olivas
Dep. of Computer Science,
University of Castilla-La Mancha,
Paseo de la Universidad 4,
13071-Ciudad Real, SPAIN
JoseAngel.Olivas@uclm.es

Pablo J. Garcés
Dep. of Computer Science,
University of Alicante
Ctra. S. Vicente del Raspeig s/n
03080 – Alicante, SPAIN
pgarces@dccia.ua.es

Abstract

In this paper it is described the use of Fuzzy Deformable Prototypes for representing clusters of documents. We use fuzzy logic technologies and a KDD based process for automatic classification of large repositories of documents in a fuzzy and hierarchical organization. The aim is to make an optimum exploitation of the concepts contained in the documents using an updateable and understandable structure. This structure allows information to be easily examined, browsed and accessed. Therefore, we apply Fuzzy Deformable Prototypes for knowledge representation. Its use allows an easy process to deal with the incoming documents and an efficient update of the structure.

Keywords: Fuzzy Prototypes, Repositories of Documents, Clustering, Knowledge Representation.

1 Introduction

The amount of available digital text documents has been growing during the last decades due to the development of new techniques for the production, storage and exchange of information. A system that could automatically organize messages would be useful for financial and world news applications

where decision making processes are based on new events and the evolution of ongoing events. Effective knowledge management is a major competitive advantage in today's information society.

Document classification or text categorization (as used in information retrieval context) is the process of assigning a document to a predefined set of categories based on the document content. However, the predefined categories are unknown in a real repository of documents. Text clustering methods can be applied to structure the resulting set of documents, so they can be interactively browsed by the user. Therefore, using a clustering process, it is possible to achieve the splitting up of the collection of documents in a reduced number of groups made up of documents with enough conceptual similarity.

The amount of documents on a repository can range from tens to thousands and the user should manage their contents efficiently. Thus, the most important features of the organization of the repository and its components (classifiers, etc.) must be the following:

- *Dimensionality:* The classifier can handle feature spaces of tens of thousands dimensions, this requires the ability to deal with sparse data spaces or a method of dimensionality reduction.
- *Efficiency:* The documental clustering algorithms must be very efficient and scalable. The method should be also accurate in the task of classifying an incoming document.

- *Understandability*: The method must provide an understandable description of the discovered clusters.
- *Updatability*: The classifier update itself promptly as each new document is filed in the repository. Also, the user can easily update the representation of the clusters of documents.

In this paper, we present, in the context of the intelligent Information Retrieval, a soft-computing based methodology that enables the efficient management of a repository of documents. In this work, the use of Fuzzy Deformable Prototypes [1], with the aim of representing the document clusters, allows the understandability and the updatability of the model. Therefore, the clusters representation could guide the user in the process of browsing through the structure.

This approach could be more representative than the standard ones that do not provide an understandable description of the documents grouped in some clusters. The fuzzy hierarchical structure and the representation of fuzzy prototypes are more realistic than the complex probabilistic models or the large amount of rules needed in other approaches. The hierarchical feature allows information to be examined and browsed at various concepts specificities, and the fuzzy organization allows information to be accessed from all related concepts.

The rest of the paper is organized as follows: In Section 2, we present the automatic process for obtaining a fuzzy hierarchical organization. The proposal of a document cluster representation based on fuzzy deformable prototypes is explained in Section 3. We describe an example in Section 4 and finally, we conclude this work in Section 5.

2 Automatic organization of documents

The construction process is based on the following stages: linguistic pre-processing, conceptual representation, and clustering.

2.1 Linguistic pre-processing

All methods of text classification require several steps of preprocessing of the data [2]. This stage extracts individual words from a document. It consists of the following steps: First, any non-textual information is removed from the documents (lexical analysis). Then, stop words such as “I”, “am”, “and”

etc. are also removed. A term is any sequence of characters separated from other terms by some delimiter. Note that a term may either be a single word or consist of several words. Typically, the terms are reduced to their basic stem applying a stemming algorithm. A candidate keyword had to appear in at least three documents and in no more than the 50% of all documents. Only the candidate keywords are useful for the following stages.

2.2 Conceptual Representation

To get a logical representation is essential to be able to work with the documents in an abstract way. For this aim, we use the FIS-CRM model (Fuzzy Interrelations and Synonymy based Concept Representation Model) [3] to extract and represent the concepts contained in the documents. This model considers the fuzzy synonymy and the fuzzy generality interrelations as a way of representing word interrelations to compound the concept definition.

The fundamental basis of FIS-CRM is to “share” the occurrences of a contained word among the fuzzy synonyms that represent the same concept and to “give” a fuzzy weight to the words that represent a more general concept than the contained one.

In this work, this model is supported by the OMCSNet Tool (A Common Sense Inference Toolkit) [4]. This tool is presently a semantic network of 280,000 items of common-sense knowledge, and a set of tools for making inferences using this knowledge. We build the fuzzy dictionaries using the predicates contained in this tool.

2.3 Clustering.

In this work, a hierarchical fuzzy clustering approach is presented. In a real context, the spaces of documents specific-to-general hierarchy and a document may belong to more than one in the hierarchy.

The clustering procedure is implemented by two connected and adapted algorithms. It uses a fuzzy hierarchical clustering algorithm to determine an initial clustering. It is well known that standard hierarchical clustering algorithms complexity is high, and increases with the number of documents. This is why the proposed algorithm is completed using the SISC [5] clustering algorithm used in FISS

meta-searcher structure [3]. The resulting organization is hierarchical, so, from a large repository of documents we will obtain a tree folders organization. The resulting clusters can be considered as fuzzy sets, so each one of the retrieved documents has a membership degree (obtained from an average similarity degree) to each one of these clusters.

3 Fuzzy Deformable Prototypes and Document Clusters Representation

The last step will consist of the analysis of the clustering results and the representation of document clusters using Fuzzy Deformable Prototypes.

3.1 Building the Fuzzy Deformable Prototypes

We represent a document cluster with a prototype, but not a classic prototype. The classic prototype theory (from psychology and also from fuzzy theory) uses a single representation of the concept of prototype, but this representation excessively simplifies the representation of the elements (documents clusters in this case). For this reason in this work it is used the concept of Fuzzy Deformable Prototype [1] based in [6]. Zadeh's idea suggests that a concept can encompass a set of prototypes, which represent the high, medium, or low compatibility of the instances with the concept. The aim must be to generate conceptual prototypes (Zadeh's approach: fuzzy schemas) that allow us to evaluate new situations from these patterns, and to establish predictions.

The use of fuzzy schemas allows us to achieve better and more understandable results, concerning patterns and prediction results. The mainly properties of the fuzzy prototypes are the following:

1. Fuzzy Deformable Prototypes can be represented as fuzzy sets. It means that it is possible to calculate a degree of "compatibility" of an element with the fuzzy set.
2. Fuzzy prototypes can be deformed [1] for describing exactly a new situation. Usually new situations must be assimilated to standard patterns.

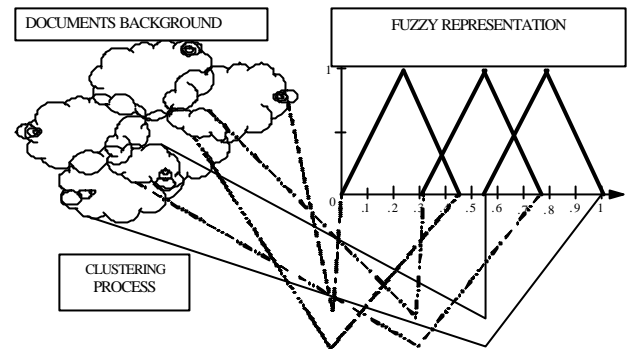


Figure 1: From Clustering to Fuzzy Prototypes

Therefore, our document clusters representation has the following components:

1. *Linguistic label*: User identification of the cluster (optional).
2. *Relevant Terms* extracted with the algorithm defined in [8]. Frequent terms set are sets of terms co-occurring in more than a threshold percentage of all documents of a database or all documents of a cluster.

$$r_i = \sum x_{ij} * d_j$$

Where

- r_i : represents the relevance of the term i in the cluster.
 - x_{ij} : represents the weight of the term i in the document j .
 - d_j : represents the membership of the document j with the cluster.
3. *Fuzzy Numbers*: The fuzzy prototypes are represented as fuzzy numbers. We obtain the center of each prototype with the following steps:

- a. Using a post-classification algorithm we split the cluster in some groups of documents or prototypes (high, medium, low affinity with the cluster).
- b. We obtain the formal definition using triangular fuzzy numbers, first normalizing the r coefficients with the following

formula:
$$r'_n = \frac{r_n - r_{\min}}{r_{\max} - r_{\min}}$$
. And last

aggregating these values:

$$c_p = \frac{\sum r_{ni} * x_i}{|R|}$$

where

- c_p is the center of the prototype.
- R is obtained aggregating the relevance of all terms in the document cluster.

The result is a representation of the clusters of documents with triangular fuzzy numbers (Fig.2).

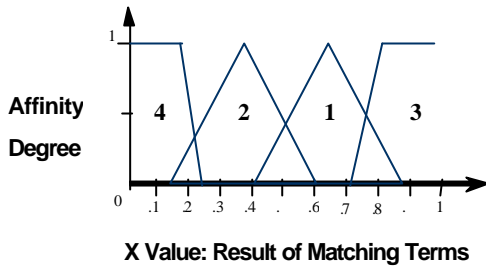


Figure 2: Fuzzy Numbers.

4. *Action Framework*: We use the framework to the parametric definition of the prototypes using standard aggregation functions. In this case, we define a framework with the probability of actions (emphasize, reject, notification, etc.).
5. *Crisp/Fuzzy Rules*: Defined ad-hoc by the users of the repository and based on the degree of membership of a new document.

The result consists on some fuzzy hierarchical groups (with special features) of documents where each document has a degree of membership to the groups in those which be integrated, and some actions associated with each group. This approach results in a small classification error when the Fuzzy Deformable Prototypes are later used for classifying new documents.

3.2 Using the Fuzzy Deformable Prototypes

The task of automatically and correctly classify each new document is complex. It is essential that the analysis and sorting operations were clear to the user, who only has to be aware of the secured results, without delays or loss of effectiveness.

The process to deal with each of the incoming documents is the following:

1. Linguistic pre-process of the message: stop words, stemming, etc.
2. Conceptual representation of the document using FIS-CRM techniques and the matrices calculated in the previous process.

3. Comparison between the characteristics of the document and the characteristics of each folder - conceptual matching based in [8]. To calculate the connection to each folder, it is used inference with Fuzzy Deformable Prototypes [1]):

- a. Obtain the X value using a distance function between vectors.
- b. Modify the X value with the fuzzy or crisp definition of user rules.
- c. Obtain the affinity of the X value with each one of the prototypes.
- d. Determine the current prototype with the modification of the prototypes, with a linear combination using a degree of affinity with the prototypes as weight values.

$$C_p(w_1...w_n) = \sum \mu_{pi}(v_1...v_n)$$

Where

- C_p : Current Prototype.
 - $(w_1... w_n)$ Parameters that describe the framework corresponding to the current prototype.
 - μ_{pi} : Affinity degree with the defined fuzzy prototypes.
 - $(v_1... v_n)$: Parameters that describe the framework corresponding to each previously defined prototype.
4. Store the message in those folders in which has been reached a positive relation. Perform the actions of the resultant framework.
 5. Update the model if the amount of incoming documents or the state of the repository requires it. This stage consists of modifying the center of the prototypes and/or the framework. Its representation eases the update of relevant coefficients.

The update of the structure could be also accomplished by a user order due to the disorganization of the repository or changes in the user's preference criteria. Another option for updating the structure is the periodical execution of a batch clustering process. Therefore, the structure

can be re-built reapplying the clustering process and reusing the previous organization.

4 Practical Example.

Some experiments have been performed to evaluate the proposed method. To evaluate the effectiveness of a classification and representation method, a test collection is needed. A test collection is an experimental tool to understand, compare and reproduce the results. Therefore, we use, in this initial experiment, the MEDLARS collection, with 1033 abstracts of medical articles. The technical note of the experiment is shown in Table 1.

Table 1: Experiment Results

| | | |
|--|--|---------------------------|
| Total documents | 750 documents 33 Clusters 15% similarity. 33% fuzzy distribution. | |
| Cluster (example) | Documents | Similarity Average |
| <i>Metabolism, manganese paramagnetic</i> | 7 | 26,95238 |
| <i>Arginine, deprivation, triglyceride</i> | 23 | 21,47036 |
| <i>Tumour, chambers, tumor-resistant</i> | 14 | 21,43956 |
| <i>Septal, angiocardigraphy</i> | 4 | 21 |
| <i>Tritiated, paraenchyma, cytidine</i> | 22 | 19,63636 |
| <i>Optokinetic, occipital, lesions</i> | 12 | 19,18182 |
| <i>Megakaryocytes, mears</i> | 11 | 18,41818 |
| <i>Regurgitation, plateaus</i> | 27 | 18,12536 |
| <i>Pneumonia, pharyngitis</i> | 12 | 18,12121 |

An example of fuzzy numbers (prototypes) of the clusters labelled with the words: “arginine, deprivation, triglyceride” (most relevant terms), are the following (Fig. 3). We have also shown in the figure the X value and the affinity degrees with the prototypes of a new document:

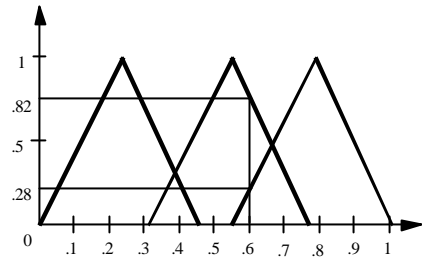


Figure 3: Affinity with the Fuzzy Prototypes.

5 Conclusions and Future Work.

First, we showed the necessity and also the problems of the classic representation of document clusters for the daily use for real documents repository user’s. Next, we presented a soft-computing based process for automatic classification of large amounts of documents in a fuzzy-hierarchical structure. Finally, we introduced the use of fuzzy deformable prototypes for the representation of clusters of documents.

The proposed method provides an understandable representation of the discovered clusters by their relevant terms and the Fuzzy Deformable Prototypes. The proposed algorithms obtain comparable results (on quality and efficiency) than the standard similar ones.

The ongoing research includes the following main issues:

- Increase the fuzzy relationships used in the conceptual representation of the documents (specialization, context, instrument, part, patient, location and agent).
- Improve the quality of frameworks of the Fuzzy Prototypes. We will build the frameworks using a co-training process [9]. This algorithm uses unlabeled data along with a few labelled examples to boost the performance of a classifier.
- Validate completely the method using different data sets which are widely used in the literature and reflect the conditions in a broad range of real life applications: MEDLARS, CISI, CACM, CRAN, etc.
- Improving the efficiency of the clustering algorithm using another approach such as Kohonen Networks or modified Fuzzy C-Means.

Acknowledgments

Partially supported by PBC-03-004 PREDACOM project, JCCM, and TIC2003-08807-C02-02 DIMOCLUST project, MCYT, Spain.

References

- [1] Olivas, J.A.: Contribution to the experimental study of the prediction based on Fuzzy Deformable Categories. PhD Thesis, University of Castilla-La Mancha, Spain (2000)
- [2] Frakes, W. B., Baeza-Yates, R.: Information Retrieval: Data Structures & Algorithms. Prentice Hall. Englewood Clifss, N. J. (1992)
- [3] Olivas, J.A., Garcés, P., Romero, F.P.: An application of the FIS-CRM model to the FISS metasearcher: Using fuzzy synonymy and fuzzy generality for representing concepts in documents. *Int. Jour. of Approx. Reasoning (Soft Computing in Recognition and Search)* (2003)
- [4] Liu, H., Singh, P.: OMCSNet v1.2. Knowledge Base, tools, and API available at: web.media.mit.edu/~hugo/ (2003)
- [5] King-ip, L., Ravikumar, K.: A similarity-based soft clustering algorithm for documents. *Proc. of the Seventh Int. Conf. on Database Sys. for Advanced Applications* (2001)
- [6] Zadeh, L. A.: A note on prototype set theory and fuzzy sets. *Cognition*, 12 (1982) 291-297
- [7] Beil, F., Ester, M., Xu X.: Frequent Term-Based Clustering. *Proceedings of the SIGKDD'02*, Edmonton, Canada (2002)
- [8] Takagi, T., Tajima, M., Proposal of a search engine based on conceptual matching of text notes, *Proc. of the BISC Int. Workshop on Fuzzy Logic and the Internet* 53-58 (2001)
- [9] Nigam, K., Ghani, R.: Analyzing the Effectiveness and Applicability of Co-training. In *Proc. of the 9th International Conference on Information Knowledge Management*, pages 86-93, McLean, VA , USA, (2000)