Document Retrieval Using Fuzzy Matching and Aggregation

Ronald R. Yager
Machine Intelligence Institute
Iona College
New Rochelle, NY 10801

A basic idea inherent in the use of fuzzy sets technology is the concept of *computing with words*. This ability will enable developers of future information retrieval systems to provide more intelligent and human friendly systems by allowing uses to interact with these systems in natural language while at the same time providing a machinary which will allow the computer to perform the types of formal operations needed. The central component of any document retrieval system is a library of documents which can be text, audio, image, video or any combination. In order to retrieve documents from this library some matching mechanism is required. To facilitate this matching process we must introduce a collection of features to help distinguish the objects in the library. Using these features each document in the library can be identified by an index consisting of values of the features associated with the document. A search is initiated by introducing a **probe** consisting of values for the features describing the users interest. The feature values in the probe are matched with the corresponding features of the objects in the library. Those documents which score the highest matching with the probe are then selected. In attempting to describe features associated with a particular environment considerable use can be made of the fuzzy set technology in that it provides a formal mechanism to enable us to more naturally represent the imprecision and gradedness associated with the concepts used in feature description.

More formally we let $V = (V_1, V_2, \ldots V_n)$ be a collection of attributes (features) which are used to index the documents and the probe. For any arbitrary object in the library we let $d = [d_1, \ldots d_n]$ be the values of the index attributes and let $p = [p_1, \ldots p_n]$ be the values of these attributes for the probe being used search the library.

The process of selecting the relevant documents from the library can be seen to consist of essentially two steps. In the first step the individual features or attributes of a document are matched to the corresponding attribute in the probe. This step results, for the document being matched, in a collection $[m_1, \ldots m_n]$, where $m_j \in [0, 1]$ is a measure of the compatibility of attribute value d_j with the probe value for that attribute, p_j. This step will be denoted as individual feature matching. In this step we can make considerable use of the facility of fuzzy sets to provide a semantics for various concepts, represent words in formal manner. In addition this formal representation brings with it a toolbox of techniques for associating degrees of matching concepts which allows for gradedness in matching process. Among these tools are measures of similarity, possibility and certainty.

The second step is the aggregation of these individual scores to obtain an overall matching value of the document to the probe. These overall scores are then used to select the relevant documents. This second step is called score aggregation. Using aggregation techniques based on fuzzy logic we can help provide implementation of many different kinds of aggregation imperatives. For example, with the aid of the ordered weighted averaging (OWA) operator we are able to model situations in which we desire that *most* of a collection of list features are satisfied. We are also able to introduce priorities as well as importances in the relationship between desired features

Our work has focused on the the development of a number of approaches which can be used by information retrieval systems builders for the implementation of these two steps. In the following we provide a brief list of references pointing to works related to this goal.

[1]. Yager, R. R. and Filev, D. P., Essentials of Fuzzy Modeling and Control, John Wiley: New York, 1994.
[2]. Yager, R. R., "On ordered weighted averaging aggregation operators in multi-criteria decision making," IEEE Transactions on Systems, Man and Cybernetics 18, 183-190, 1988.
[3]. Yager, R. R., "Connectives and quantifiers in fuzzy sets," Fuzzy Sets and Systems 40, 39-76, 1991.
[4]. Yager, R. R., "A note on weighted queries in information retrieval systems," J. of the American Society of Information Sciences 38, 23-24, 1987.
[5]. Yager, R. R., "A logical bibliographic searcher: An application of fuzzy sets," IEEE Trans. on Systems, Man and Cybernetics 10, 51-53, 1980.