Fuzzy Logic for Measuring Information Retrieval Effectiveness

Abstract: We present a new fuzzy extension of the classical effectiveness measures of information retrieval by a new way to calculate the relative cardinality of a fuzzy set. Previous approaches using Zadeh’s cardinality are compared to our new approach in an experimental stage. The experiments have been carried out with a genetic algorithm where the fitness function to optimize is a combination of the fuzzy recall and fuzzy precision measures. Results included at the end of the paper show the goodness of our proposal.

1. Introduction

Fuzzy logic has been widely utilized in different stages of the retrieval process. The representation of terms in documents and the queries with fuzzy grade in the matching task has generated the use of effectiveness measures using fuzzy logic.

The extension of the classical measures to the fuzzy case has been due to two main reasons:

- On the one hand, this extension comes from the handling of fuzzy value in the document representation by terms, since the cardinal calculus must be carried out with fuzzy sets. Therefore, the use of fuzzy cardinals is fundamental to deal with this type of representation.
- On the other hand, measures based on fuzzy logic give the flexibility that classical measures lack, since retrieved documents are counted as relevant or non-relevant without any grade in the classical case.

As is well known, the most used measures in the information retrieval framework are the recall and precision. Since the first extension of these measures to the fuzzy case in 1981, there have been several approaches in the literature in order to improve the behaviour of these measures to be used in Fuzzy Information Retrieval Systems (Gedeon and Koczy, 1995), (Sanchez y Pierre, 1994), (Martin-Bautista, 2000). These approaches can be mainly differentiated by the way the extension of the classical measures to their fuzzy extension is proposed. Concretely, the two first approaches, together with the original proposal from Kraft and Buell are based in the use of the Zadeh’s relative fuzzy cardinality to determine the number of documents belonging to a certain set in terms of relevance. However, in the third proposal, a different approach using a new way to calculate the fuzzy cardinality presented by Delgado, Sánchez y Vila en (Delgado et al., 2000) is used. We present in this work a new extension of the effectiveness measures for information retrieval based on fuzzy logic using this cardinality, which is based on the evaluation of quantified sentences.
2. Information Retrieval Measures with Fuzzy Logic

The performance criterion considered in the evaluation of an information retrieval system is that the query answer retrieved by the system should correspond to the user preferences, that is, the documents should be evaluated with the same grade of relevance that the user would do with all the documents in the collection.

The extension of the relevance-based effectiveness measures (Salton and McGill, 1983) consists of the transformation of the cardinalities of the sets into fuzzy cardinalities. The recall $\rho$ and precision $\psi$ measures in the crisp case are defined as shown in (1), where $\Omega_R$ and $\Omega_L$ represent the subsets (of $\Omega=\{\omega_1, K, \omega_n\}$) of the documents retrieved and the relevant documents, respectively.

$$\rho = \frac{|\Omega_R \cap \Omega_L|}{\Omega_L} \quad \psi = \frac{|\Omega_R \cap \Omega_L|}{\Omega_R} \quad (1)$$

Let $\Omega=\{\omega_1, ..., \omega_n\}$ be the document set queried by a query $Q$. The user evaluation of $\Omega$ with respect to $Q$ is characterized by a fuzzy subset of $\Omega$, $E_{\Omega,Q} = \beta_1/\omega_1 + \ldots + \beta_n/\omega_n$, where $\beta_i = \mu^Q_\omega(\omega_i)$ is the user’s evaluation of the degree to which the object $\omega_i$ satisfies the query $Q$. The system answer to $Q$ over $\Omega$ is also modeled by a fuzzy subset of $\Omega$; let us denote this subset $S_{\Omega,Q} = \alpha_1/\omega_1 + \ldots + \alpha_n/\omega_n$.

In order to get a perfect information retrieval system, the ranking given by the system must be the same as the ranking that the user gives to the same set of documents, that is, $S^{(\alpha)}_{\alpha_i} = E^{(\beta)}_{\beta_i}$ for all $i$, where $S^{(\alpha)}_{\alpha_i}$ and $E^{(\beta)}_{\beta_i}$ are respectively the $\alpha_i$-cut of $S_{\Omega,Q}$ and the $\beta_i$-cut of $E_{\Omega,Q}$.

The fuzzy extension of (1), applying the sigma count as the (scalar) cardinality (Zadeh, 1983), is given by:

$$\tilde{\rho} = \frac{\sum_{\alpha_i \in \Omega} \mu^S_{\alpha_i}(\omega_i) \wedge \mu^E_{\alpha_i}(\omega_i)}{\sum_{\alpha_i \in \Omega} \mu^E_{\alpha_i}(\omega_i)} \quad \tilde{\psi} = \frac{\sum_{\alpha_i \in \Omega} \mu^S_{\alpha_i}(\omega_i) \wedge \mu^E_{\alpha_i}(\omega_i)}{\sum_{\alpha_i \in \Omega} \mu^S_{\alpha_i}(\omega_i)} \quad (2)$$

Several approaches have been presented in the literature concerning the use of fuzzy logic in information retrieval. As we commented above, a fuzzy representation of the terms is needed for these fuzzy models. Such a representation depends on the retrieval system considered. In the following, we explain the main approaches to fuzzy measures.

2.1. Buell and Kraft approach

These authors made one of the first extensions of Boolean retrieval systems with the fuzzy model (Buell and Kraft, 1981). Their fuzzy representation was based in a fuzzy membership function of “aboutness“ of a term to a document. The performance measurement of the model is a generalization of the recall and precision measures, calculated from two rankings of evaluations $S$ and $E$ in an analogous manner as the crisp case:

$$\rho = \frac{|S \cap E|}{|E|} \quad \psi = \frac{|S \cap E|}{|S|} \quad (3)$$
where \( S \) and \( E \) may be interpreted as the system's and expert's evaluation, respectively.

Using Zadeh's cardinality, these expressions are transformed into the fuzzy equivalent ones:

\[
\tilde{\rho} = \frac{\sum_{i=1}^{n} (\beta_i \land \alpha_i)}{\sum_{i=1}^{n} \beta_i} \quad \tilde{\psi} = \frac{\sum_{i=1}^{n} (\beta_i \land \alpha_i)}{\sum_{i=1}^{n} \alpha_i}
\]  

(4)

2.2. Sanchez and Pierre approach

In this approach (Sanchez and Pierre, 1994), a fuzzy extension of the term weighting is considered for Boolean retrieval systems. In this model, terms have two fuzzy values associated. On the one hand, the value of "aboutness", meaning the degree in which the term is related to a document. On the other hand, a fuzzy query weight is associated to a term and coded into chromosomes. This weight means the relative importance of the term in the query. The fitness function to evaluate is a combination of fuzzy precision and fuzzy recall. For the output, documents are retrieved according to a threshold \( T \) in answer to the query considered. The fuzzy recall and precision measures are defined basing on the extension from cardinalities to fuzzy cardinalities of the classical measures, as is presented in (5).

\[
\tilde{\rho} = \frac{\sum_{\alpha_i \in \Omega} \mu_{\Theta_a} (\alpha_i) \land \mu_{\Theta_k} (\alpha_i)}{\sum_{\alpha_i \in \Omega} \mu_{\Theta_k} (\alpha_i)} \quad \tilde{\psi} = \frac{\sum_{\alpha_i \in \Omega} \mu_{\Theta_a} (\alpha_i) \land \mu_{\Theta_k} (\alpha_i)}{|C^\theta_R|}
\]  

(5)

Although these two approaches are based on Zadeh's cardinality, there is a difference between them. In the former, all the fuzzy values of the documents in a ranking are considered. In the latter, a relevance threshold \( \theta \) is fixed, but the cardinality of the denominator in the case of the fuzzy precision is calculated as the cardinality of a crisp set of the retrieved documents above the value \( \theta \).

The expressions of recall and precision (4) measure the relative cardinality of two sets. The relative cardinality of one set \( A \) with respect to another set \( B \) defined on the same domain is the percentage of objects of \( B \) that are in \( A \). This concept has been generalized to the fuzzy case in several ways. The most used measure is Zadeh's (Zadeh 1975), which is based on the \( \Sigma \)-count(F) measure introduced by (De Luca and Termini, 1972). The \( \Sigma \)-counts measure the energy of a fuzzy set (i.e., the total amount of membership to the set), rather than the integer number of objects in the fuzzy set (De Luca and Termini 1972). In fact, it is known that sometimes \( \Sigma \)-counts provide a high value for the cardinality of a fuzzy set due to an accumulation of small membership degrees. This is counterintuitive. When measuring relative cardinalities the situation can be worst. More details are given in (Delgado et al., 1999), where it is claimed that the only possible cardinalities of a fuzzy set are the cardinalities of its \( \alpha \)-cuts.
3. Fuzzy Measures with Delgado, Sánchez and Vila fuzzy relative cardinality

The extension of the measures presented is based on the definition of new measures of the fuzzy relative cardinality of fuzzy sets (Delgado et al. 1999) and the evaluation of type II quantified sentences (Zadeh 1983) by means of a new method that was introduced in (Delgado et al. 2000). This last approach has been successfully employed in data mining applications to obtain conditional evidences involving fuzzy items (Sánchez 1999).

3.1. Fuzzy Recall

The expression of fuzzy recall with the mentioned cardinality calculus is as follows:

\[
\rho = DSV(S/E) = \sum_{\alpha_i \in \Delta(S/E)} (\alpha_i - \alpha_{i+1}) \frac{(S \cap E)^{[\alpha_i]}}{[\alpha_i]}
\]

where \(\Delta(S/E) = \Lambda(S \cap E) \cup \Lambda(E) = \{\alpha_1, \ldots, \alpha_{2n}\} \) being the cardinality of the set \(\Delta(S/E)\), and with \(1 = \alpha_1 > \alpha_2 > \ldots > \alpha_{q+1} = 0\), \(\Lambda(F) = \{F(x) \forall x \in X, F(x) > 0\}\) being the set of levels of \(F\), \(F(\alpha)\) being the \(\alpha\)-cut of \(F\), and \(\alpha_i\) being the classical set cardinality. The set \(E\) is assumed to be normalized before the recall. If not, \(E\) is normalized (i.e. each degree is divided by a factor equal to \(\max\{E(x) \mid x \in X\}\)) before the precision is calculated, and the same normalization factor is applied to \(S \cap E\).

3.2. Precision

In an analogous way, the precision is calculated as in (6).

\[
\psi = DSV(E/S) = \sum_{\alpha_i \in \Delta(E/S)} (\alpha_i - \alpha_{i+1}) \frac{(S \cap E)^{[\alpha_i]}}{[\alpha_i]}
\]

where \(\Delta(E/S) = \Lambda(S \cap E) \cup \Lambda(E) = \{\alpha_1, \ldots, \alpha_p\} \) being the cardinality of the set \(\Delta(E/S)\), and with \(1 = \alpha_1 > \alpha_2 > \ldots > \alpha_{p+1} = 0\). In this case, the set \(S\) is assumed to be normalized. If not, \(S\) is normalized before the precision is calculated, and normalization factor is applied to \(S \cap E\).

Both expressions (6) and (7) can be obtained in \(O(n)\) (\(n\) being the number of documents) if we consider only a fixed set of levels \(\Delta(E/S)\), (Sánchez, 1999).

4. Experimental Stage

4.1. Description of the System

For the experiments, we consider the genetic model oriented to documents presented in (Martín-Bautista, 2000), where we start from a set of documents \(\Omega = \{D_1, \ldots, D_K, D_{K+1}, \ldots, D_m\}\) ordered decreasingly and evaluated by the user, being \(u_i \in [0,1]\) the user’s evaluation of the document \(D_i, i = 1, \ldots, m\), that is, the degree that the user consider a document relevant. The set \(\Omega\) is divided into
two subsets $\Omega_R = \{D_1, \ldots, D_K\}$ and $\Omega_{NR} = \{D_{K+1}, \ldots, D_m\}$, containing the relevant documents and the non-relevant documents, respectively.

The relevance threshold from which we consider a document relevant is $\alpha \in [0,1]$, initially fixed to 0.5. We shall assume that an evaluation $\nu > 0.5$ indicates a good document, while $\nu \leq 0.5$ indicates a bad document. The fitness function to optimize is a combination of fuzzy recall and fuzzy precision as follows, called fuzzy recall-precision ($\tau$) given by:

$$\tau = \rho^\nu \psi^{\nu_2}$$  \hspace{1cm} (8)

where $\rho$ is the fuzzy recall, and $\psi$ is the fuzzy precision defined in the former section and $\nu_1, \nu_2$ are the importance weights of fuzzy recall and fuzzy precision, respectively.

Let $T = \{t_1, Kt_n\}$ be the set of terms extracted from document base $\Omega$, and $x_{ij}$ the relative (normalized) frequency of term $t_j$ in document $D_i$. The estimation of the expected value of $x_j$ in good and bad documents is given by the weighed average of the relative occurrence frequency of a symbol $t_j$ in the collection. The knowledge of the system about the user preferences is kept in the population of a Genetic Algorithm. A gene in a chromosome is defined by a term and a fuzzy number $\widetilde{n}$ of occurrences of the term in documents belonging to the class of documents that satisfy the user's information need. The details of this fuzzy representation can be seen in (Martin-Bautista, 2000).

4.1.1. Genetic Components

The genetic components have been determined by preliminary tests. The selection scheme for probabilities is the inverted linear ordering (Bäck, 1992). The selection mechanism is the universal stochastic sample (Goldberg, 1989), with an elitism model. To carry out the crossover, we have considered the one-point crossover operator, with a probability of 0.6. As for the mutation, we choose by a standard mechanism a gene to mutate. Finally, the size of the population is 80 chromosomes and the chromosome length is 10. We have considered 1000 generations for each run of the algorithm, calculating the average of three runs with different random seeds to get the final results. The importance weights of the fuzzy recall ($\nu_1$) and fuzzy precision ($\nu_2$) are set as $\nu_1 = 0.67$ and $\nu_2 = 1$.

4.2. Document Sets

We have considered two different examples to carry out the experimental stage. The first example is a collection of documents corresponding to the follow query in the INSPEC database of Jul-Sep 1998: "Information Retrieval and Classification". The number of documents retrieved was 22, and the number of different terms extracted (after removing stop-list words and stemming) is 616. The number of total terms is 1,074. Let suppose that the user's information needs are oriented to those documents (in the collection of the 22 documents retrieved previously), regarding topics such as Web and Internet, to which the user will give the higher evaluation. Initially, the documents are evaluated by an expert, where there are 8 relevant documents and 14 non-relevant ones, with the relevance threshold set to 0.5.

The second example is a collection of 100 documents, corresponding to the query "Information and Retrieval" to the INSPEC database of 1999-2000. The number of terms extracted for these documents is 11,210, where 1641 terms are
different. The preferences of the user in the framework of this query are located in the field of Genetic Algorithms. As in the first example, the user will give the highest relevance to the most preferred documents. In this case, the number of relevant documents is 13 and there are 87 non-relevant documents.

4.3. Classification Errors

Additionally to the effectiveness information retrieval measures, we can define classification errors to check the goodness of the measures, and compare among the fuzzy measures using Zadeh’s relative cardinality and Delgado, Sánchez and Vila’s We can compute a global error comparing the relevance that the system gives to the documents to the relevance that user would give (total error). Moreover, we can compute the partial errors taking into account on the one hand the relevant documents (relevant error) and the non-relevant documents (non-relevant error).

\[
\text{Total error} = \frac{\sum_{i=1}^{[\Omega]} (t_i - s_i)^2}{|\Omega|} \quad \text{Relevant error} = \frac{\sum_{i=1}^{[\Omega]} (t_i - s_i)^2}{|\Omega_r|} \quad \text{Non relevant error} = \frac{\sum_{i=1}^{[\Omega]} (t_i - s_i)^2}{|\Omega_{nr}|}
\]

4.4. Results

As we can see in the tables below, the experiments has been realized with the two examples explained in section 4.2. In Table 1, the values of fuzzy recall-precision, fuzzy recall and fuzzy precision measures are shown using Zadeh’s and DSV’s cardinality. As can be observed this table, the difference between the values of the measures for both cardinalities can be noted, specially in the example 2, when the number of documents is higher (100). More exactly, the values of the measures for DSV’s cardinality are lower than the ones obtained with Zadeh’s cardinality. This fact would lead us to think that one of the measures is not as exact as we expect.

<table>
<thead>
<tr>
<th>Example</th>
<th>Cardinality</th>
<th>Fuzzy Recall - Precision</th>
<th>Fuzzy Recall</th>
<th>Fuzzy Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Zadeh</td>
<td>0.395</td>
<td>0.632</td>
<td>0.537</td>
</tr>
<tr>
<td></td>
<td>DSV</td>
<td>0.333</td>
<td>0.581</td>
<td>0.467</td>
</tr>
<tr>
<td>2</td>
<td>Zadeh</td>
<td>0.256</td>
<td>0.545</td>
<td>0.388</td>
</tr>
<tr>
<td></td>
<td>DSV</td>
<td>0.185</td>
<td>0.380</td>
<td>0.356</td>
</tr>
</tbody>
</table>

Table 1. Effectiveness fuzzy measures with different cardinalities

Let us observe Table 2, where the classification errors of the same examples with both cardinalities are calculated. In example 1, both errors are almost equal, although results with the DSV’s cardinality seem slightly higher. However, in example 2, the error for Zadeh’s cardinality is higher than DSV’s one, specially in the non relevant error.

This fact corroborates our assessment that Zadeh’s cardinality provides a high value for the cardinality of a fuzzy set due to an accumulation of small values.
5. Concluding Remarks and Future Work

In this work, we have presented new effectiveness information retrieval measures with fuzzy logic using alternatives to Zadeh's cardinality. It has been shown that DSV cardinality provides more intuitive cardinalities with an efficiency of $O(n)$, with $n$ being the number of documents. Results have shown that the classification error is higher when the number of documents increase for measures with Zadeh's cardinality, while the values for the fuzzy recall and fuzzy precision measures is higher than in measures with the DSV's measures, which is contradictory.

In a future work, a deeply study of the experimental cases with different sets of documents with respect to their size and the number of relevant and non relevant documents within them will be presented.

<table>
<thead>
<tr>
<th>Example</th>
<th>Cardinality</th>
<th>Total error</th>
<th>Relevant error</th>
<th>Non Relevant Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Training</td>
<td>Test</td>
<td>Training</td>
</tr>
<tr>
<td>1</td>
<td>Zadeh</td>
<td>0.20</td>
<td>0.20</td>
<td>0.177</td>
</tr>
<tr>
<td></td>
<td>DSV</td>
<td>0.21</td>
<td>0.20</td>
<td>0.192</td>
</tr>
<tr>
<td>2</td>
<td>Zadeh</td>
<td>0.26</td>
<td>0.26</td>
<td>0.329</td>
</tr>
<tr>
<td></td>
<td>DSV</td>
<td>0.22</td>
<td>0.22</td>
<td>0.398</td>
</tr>
</tbody>
</table>

Table 2. Classification errors with different cardinalities.

References


