FUZZY INTERVAL NUMBER (FIN) TECHNIQUES FOR MULTILINGUAL AND CROSS LANGUAGE INFORMATION RETRIEVAL

Theodoros Alevizos, Vassilis G. Kaburlasos, Stelios Papadakis
Department of Industrial Informatics, T.E.I. of Kavala, Greece

Christos Skourlas, Petros Belsis
Department of Informatics, T.E.I. of Athens, Greece

Keywords: Fuzzy Interval Number (FIN), Information Retrieval (IR), Cross Language Information Retrieval (CLIR).

Abstract: Fuzzy Interval Numbers (FINs) could be seen as a set of techniques applied in Fuzzy System applications. In this paper, we propose a series of techniques to solve multi-Lingual and Cross Language Information Retrieval (CLIR) problems, based on Fuzzy Interval Numbers (FINs). Some experiments showing the importance of these techniques in the CLIR-systems are briefly described and discussed. Our method is evaluated using monolingual and bilingual public bibliographic data extracted from the National Archive of the Greek National Documentation Centre. All the experiments were conducted with and without the use of stemming, stop-words and other language dependent (pre-) processing techniques. It seems that a main advantage of our approach is that the method is language independent and there is also no need for any text pre-processing or higher level processing, avoiding thus the use of taggers, parsers, feature selection strategies, or the use of other language dependent NLP tools.

1 INTRODUCTION

Fuzzy (set) techniques were proposed for Information Retrieval (IR) applications many years ago (Radecki, 1979), (Kraft, 1993), mainly for modelling. Fuzzy Interval Numbers (FINs) and the related theoretical (mathematical) background, which is based on the metric space of the generalized intervals, were introduced by Kaburlasos (Kaburlasos, 2004), (Petridis et al, 2003) in fuzzy system applications. A FIN (see Figure 1) may be interpreted as a conventional fuzzy set; additional interpretations for a FIN are possible including a statistical interpretation.

Fuzzy Interval Numbers (FINs) and lattice algorithms have been employed in various real-world applications including numeric and non-numeric data. Kaburlasos and Petridis (Kaburlasos et al, 2000), Petridis and Kaburlasos (Petridis et al, 1998, 2000, 2001) presented FLN (Fuzzy Lattice Neurocomputing mainly for competitive clustering. The most popular among the FLN models for clustering is σ-FLN. For example, (Petridis and Kaburlazos, 2000) describe the automated – electromechanical surgical *mechatronics tool* which is used to control penetration through soft tissues in the epidural punctury. In this framework a series of learning experiments for soft tissues recognition was carried out and the σ-FLN model and the Voting σ-FLNMAP algorithm were applied.

FLN algorithms were also used in stapedotomy surgery (Kaburlasos et al, 1997), and prediction of ozone concentration by classification, based on meteorological and air quality data (Athanasiadis and Mitkas, 2004).

(Petridis et al, 2001), (Kaburlasos et al, 2002), (Petridis and Kaburlasos, 2003) reported a best sugar prediction accuracy using Fuzzy Interval Numbers and the FINkNN classifier and a population (measurements) of production and meteorological data. (Kaburlasos et al, 2005) used FINs for representing geometric and other fertilizer granule features. This type of modelling was used to cover needs of the Greek Fertilizer Industry.

(Marinagi et al, 2006) proposed the use of FINs classifier to handle problems of Cross Language

Alevizos T., G. Kaburlasos V., Papadakis S., Skourlas C. and Belsis P. (2007).
FUZZY INTERVAL NUMBER (FIN) TECHNIQUES FOR MULTILINGUAL AND CROSS LANGUAGE INFORMATION RETRIEVAL.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - AIDSS*, pages 348-355
DOI: 10.5220/0002401503480355

Information Retrieval. The basic features of this method are the following:

1) Documents are represented as FINs (see Figure 1).

2) The FIN representation of documents is based on the use of the collection term frequency as the term identifier.

3) The use of FIN distance instead of a similarity measure.

For the distance calculations a bell-shaped mass function was used:

$$m_h(t) = \frac{\alpha + \beta h}{A^2 + \left(t - \frac{\max(ctf)}{2}\right)^2}$$

The positive real numbers A, α, β, are parameters; max(ctf) is the maximum value of all the collection term frequencies.

The structure of the remainder of this paper is as follows:

In section 2 a brief introduction to the mathematical background of the generalized intervals and other relevant concepts is given. The notation used (Kaburlasos, 2004) is an evolution (simplification) of the one described by (Marinagi et al, 2006). In section 3 the "conceptual" transition from document vectors to document FINs is presented. Section 4 presents how to calculate the similarity of documents using Fuzzy Interval Numbers. In section 5 we present our experiments and evaluate and discuss our method. Conclusions and current work on the topic are presented in section 6.

2 THEORETICAL BACKGROUND

2.1 Generalised Intervals

A positive generalized interval of height h ∈ (0,1] is a map $\mu_{[x_1,x_2]^h} : R \longrightarrow \{0, h\}$, given by:

$$\mu_{[x_1,x_2]^h}(x) = \begin{cases} h, & x_1 \le x \le x_2 \\ 0, & \text{otherwise} \end{cases}$$

where $x_1 \le x_2$

A negative generalized interval of height h(∈ (0,1] is a map $\mu_{[x_1,x_2]^h} : R \longrightarrow \{0, -h\}$, given by:

$$\mu_{[x_1,x_2]^h}(x) = \begin{cases} -h, & x_1 \ge x \ge x_2 \\ 0, & \text{otherwise} \end{cases}$$

where $x_1 > x_2$

We shall use below the more compact notation $[x_1, x_2]^h$ instead of the μ notation.

The interpretation of a generalized interval depends on an application; for instance if a feature is present it could be indicated by a positive generalized interval. Generalized intervals will be used for introducing a metric into the lattice of the Fuzzy Interval Numbers (FINs) below.

The set of all positive generalized intervals of height h is denoted by M_+^h, the set of all negative generalized intervals by M_-^h, and the set of all generalized intervals by $\mathbf{M}^h$.

2.1.1 The Basic Idea

FIN is constructed (see CALFIN algorithm below) such that any horizontal line ε_h, h∈ [0, 1], intersects a FIN at exactly two points (only for h=1 there exists a single intersection point). Hence, a horizontal line ε_h results in a "rectangular shaped pulse" of height h which is called generalized interval of height h. If a metric distance could be defined between every two generalized intervals of height h then a metric distance is implied by two FINs simple by computing the corresponding definite integral from h=0 to 1.

In figure 1 we can see two intersecting generalized intervals [a', c'], [b', d'] at the height of h. The intervals could be mapped to the conventional intervals [a, c], [b, d]. The area "under" a generalized interval is a real number which could be calculated. We can define a metric distance and an inclusion measure function in the set (lattice) of the generalized intervals $\mathbf{M}_h$ based on these notes.

2.1.2 The Lattice of All the Generalised Intervals

Two functions are defined in the set of all the generalized intervals M^h:

1. Function support maps a generalized interval to the corresponding conventional interval; support $([x_1, x_2]^h) = [x_1, x_2]$ for positive, support$([x_1, x_2]^h) = [x_2, x_1]$ for negative and support$([x_1, x_2]^h) = \{x_1\}$ for trivial generalized intervals.

2. Function sign: $\mathbf{M}^h \rightarrow \{ -1, 0, +1 \}$ maps a positive generalized interval to +1, a negative generalized interval to –1 and a trivial generalized interval to 0.

A partial ordering relation ≤ can be defined in the set $\mathbf{M}^h$, h∈ (0,1]:

1) $[a, b]^h \le [c, d]^h \Leftrightarrow$ support$([a, b]^h) \subseteq$ support $([c, d]^h)$, for $[a, b]^h$, $[c, d]^h \in \mathsf{M}_+^h$

2) $[a, b]^h \leq [c, d]^h \Leftrightarrow$ support($[c, d]^h$) $\subseteq$ support($[a, b]^h$), for $[a, b]^h$, $[c, d]^h \in \mathbf{M}_{-}^{h}$

3) $[a, b]^h \leq [c, d]^h \Leftrightarrow$ support($[c, d]^h$) $\cap$ support($[a, b]^h$) $\neq 0$, for $[a, b]^h \in \mathbf{M}_{-}^{h}$, $[c, d]^h \in \mathbf{M}_{+}^{h}$

Kaburlazos (Kaburlasos,2004) proved that the set of all generalized intervals M^h is a lattice. If q_1, $q_2 \in M^h$ are the "intersecting" positive generalized intervals $[a', c']^h$, $[b', d']^h$ which are shown in figure 1 then the join $q_1 \vee q_2$ is equal to $[a', d']^h$ and the meet $q_1 \wedge q_2$ is equal to $[b', c']^h$. In a similar way we can defined the meet and join in the case of non intersecting positive, intersecting negative, and non intersecting negative generalized intervals (Marinagi et al, 2006).

2.1.3 A Metric Distance in the Set of All Generalized Intervals M^h

A valuation v in a lattice L, defined as the area "under" a generalized interval, is a real function v: L → R which satisfies $v(x)+v(y)= v(x \vee_L y) + v(x \wedge_L y)$, x,y ∈ L.A valuation is called monotone if and only if $x \leq y$ implies $v(x) \leq v(y)$ and positive if and only if $x < y$ implies $v(x) < v(y)$ for x,y ∈ L.

The role of a positive valuation function v: L→ R is to be a mapping from a lattice L of semantics to the mathematical field R of real numbers for carrying out computations.

Therefore a metric distance d:L×L→ R can be defined in the lattice $\mathbf{M}^h$, h∈ (0,1] given by $d(x, y) = v(x \vee_L y) - v(x \wedge_L y)$, x,y ∈ L.

Kaburlazos (Kaburlasos, 2004) proved the following proposition:

Proposition

Let the underlying positive valuation function f: R→ R be a strictly increasing real function in R. Then the real function v: $\mathbf{M}^h \rightarrow R$ is given by

$$v([a, b]^h) = \text{sign}([a, b]^h)\, c(h)\, {}_a\!\int^b [f(x) - f(a)]\, dx$$

where v is a positive valuation function in $\mathbf{M}^h$, c: (0,1] → R^+ is a positive real function for normalization. A metric distance in $\mathbf{M}^h$ is given by:

$$d_h(x, y) = v(x \vee y) - v(x \wedge y)$$

Therefore, we can choose a positive function f to define the valuation function v and also simplify the calculation of the distance between the generalized intervals at a height h, $0 \leq h \leq 1$ In the following paragraphs we can see how to construct a strictly increasing real function f_h:R→R and simplify the calculation of the definite integral. We can use an integrable mass-function m_h: R→ R_0^+ as: $f_h(x) = \int_0^x m_h(t)dt$.

Various mass functions can be considered:

For mass-function $m_h(x)=h$ it follows metric,

$$d_h([a, b]^h , [c, d]^h)= h\, (|a-c|+|b-d|).$$

For mass-function $m_h(x)=3x^2$ the corresponding function is $f_h(x)=x^3$ and therefore,

$$d_h([-1,0]^1, [3,4]^1) = [f_h(-1\vee 3) - f_h(-1\wedge 3)] + [f_h(0\vee 4) - f_h(0\wedge 4)] = 92.$$

If the mass function $m_h(x)$ is equal to,

$$\frac{2e^{-x}}{(1+e^{-x})^2}$$

the corresponding (logistic) function is

$$f_h(x) = \frac{2}{1+e^{-x}} - 1 = \frac{1-e^{-x}}{1+e^{-x}}$$

and therefore,

$$d_h([-1,0]^1, [3,4]^1) = [f_h(-1\vee 3) - f_h(-1\wedge 3)] + [f_h(0\vee 4) - f_h(0\wedge 4)] = 2.3312.$$

3 FUZZY INTERVAL NUMBERS: DEFINITION AND COMPUTATION

Given a population (a vector of real numbers) $x = [x_1,x_2,\ldots,x_N]$ of measurements, sorted in ascending order, a FIN can be computed by applying the CALFIN algorithm below. FIN is regarded as an abstract "mathematical object" and could have various interpretations and uses. The notation dim(x) denotes the dimension of vector x, e.g. dim([2,-1])= 2, dim([-3,4,0,-1,7])= 5, etc. The median(x) of a vector $x = [x_1,x_2,\ldots,x_N]$ is defined to be a number such that half of the N numbers $x_1,x_2,\ldots,x_N$ are smaller than median(x) and the other half are larger than median(x); for instance, the median($[x_1,x_2,x_3]$) with $x_1 < x_2 < x_3$ equals x_2, whereas the median($[x_1,x_2,x_3,x_4]$) with $x_1 < x_2 < x_3 < x_4$ was computed here as median($[x_1,x_2,x_3,x_4]$)= $(x_2 + x_3)/2$.

Algorithm CALFIN

```
1. Let x be a vector of real numbers.
2. Order incrementally the numbers in vector x.
3. Initially vector pts is empty.
4. function calfin(x) {
5. while (dim(x) ≠ 1)
6. medi:= median(x)
7. insert medi in vector pts
8. x_left:= elements in vector x less-than number median(x)
9. x_right:= elements in vector x larger-than number median(x)
```

10. calfin(x_left)
11. calfin(x_right)
12. endwhile
13. } //function calfin(x)
14. Sort vector pts incrementally.
15. Store in vector val, dim(pts)/2 numbers from 0 up to 1 in steps of 2/dim(pts) followed by another dim(pts)/2 numbers from 1 down to 0 in steps of 2/dim(pts).

The above procedure is repeated recursively $\log_2 N$ times, until "half vectors" are computed including a single number; the latter numbers are, by definition, median numbers. The computed median values are stored (sorted) in vector pts whose entries constitute the abscissae of a positive FIN's membership function; the corresponding ordinate values are computed in vector val. Note that algorithm CALFIN produces a positive FIN with a membership function $\mu(x)$ such that $\mu(x)=1$ for exactly one number x.

Now we can focus on the transition from document vectors to document FINs

FINs and documents' representation and construction

In the Vector Space Model for Information Retrieval, a text document is represented by a vector in a space of many dimensions, one for each different term in the collection. In the simplest case, the components of each vector are the frequencies of the corresponding terms in the document:

$$Doc_k = (f_{k1}, f_{k2}, \dots f_{kn})$$

f_{kj} stands for the frequency of occurrence of term t_j in document Doc_k.

Example Table depicts the vector space model of a small collection comprising four documents:

$tf_{kj} = tf_{2,12}=6$ stands for the frequency of occurrence of term t_{12} in document Doc_2.

$ctf_j = ctf_{12}=7$ stands for the total frequency of occurrence of term t_{12} in the whole collection.

Then ctf_{12} is equal to

$$\sum_{k=1,4} tf_{k1,2} = tf_{1,12} + \ldots + tf_{4,12} = 0+6+0+1=7$$

The total frequency, of occurrence of term t_j in the whole collection, ctf_j is equal to $\sum_k tf_{kj}$

The collection term frequencies (ctf) are used as term identifiers. In order to ensure the uniqueness of the identifiers a multiple of a small ε is added to the ctfs when needed (see last column of the table).

If we want to compute the FIN of the Doc1 then we focus on the columns: Terms, Doc1, Term Identifiers. We repeat the non-zero values of keys as many times as the terms are contained in the Doc1. In order to ensure the uniqueness of the identifiers (keys) we add a multiple of a small ε to the ctfs when needed again. The FIN of the document will be computed from the identifiers of the terms that exist in the document (19 values in our case). Eventually, the abscissae vector is exactly the "number population" from which the document FIN is computed from by the CALFIN algorithm: X=(4.333,4.334,4.335,4.667,4.668, 5, 5.25, 5.251, 5.252, 5.5, 5.501, 5.75, 6, 8, 8.001, 8.5, 8.501, 8.502,9). Figure 1 shows two FINs which are calculated by our algorithm.

4 FUZZY INTERVAL NUMBERS AND DOCUMENTS' SIMILARITY

The FIN function F: (0,1] → M maps a real number h in (0,1] to a generalized interval F(h). The domain of function F is shown on the vertical axis, whereas the range of function F includes "rectangular shaped pulses" (generalized intervals at height h, $h\in[0,1]$) on the plane. Now we can use the concepts we have already seen in the case of the generalized intervals. Hence, a positive Fuzzy Interval Number (FIN) is defined as a continuous function F: $(0,1] \to \mathbf{M}_+^h$ such that:

$h_1 \leq h_2 \Rightarrow \text{support}(F(h_1)) \supseteq \text{support}(F(h_2))$, where $0 < h_1 \leq h_2 < 1$.

The set of all positive FINs is denoted by $\mathbf{F}_+$. Similarly, negative FINs are defined. The set F of FINs is partially ordered by an ordering relation $\leq_F$, which is introduced as follows:

Let $F_1, F_2 \in F$, then $F_1 \leq_F F_2$ if and only if $F_1(h) \leq_M^h F_2(h)$

The relation $\leq_F$ is reflexive, anti-symmetric and transitive. There are incomparable FINs but we can define the meet and the join, again. Using the metric distance between two generalized intervals $F_1(h)$ $(=[a,b]^h)$ and $F_2(h)$ $(=[c,d]^h)$ Kaburlasos (Kaburlasos, 2004) proved the following proposition:

Proposition

Given two positive FINs F_1 and F_2, then

$$d(F_1,F_2)= c\int_0^1 d_h(F_1(h), F_2(h))dh$$

where c is an user-defined positive constant, is a metric distance

Calculation of the distance

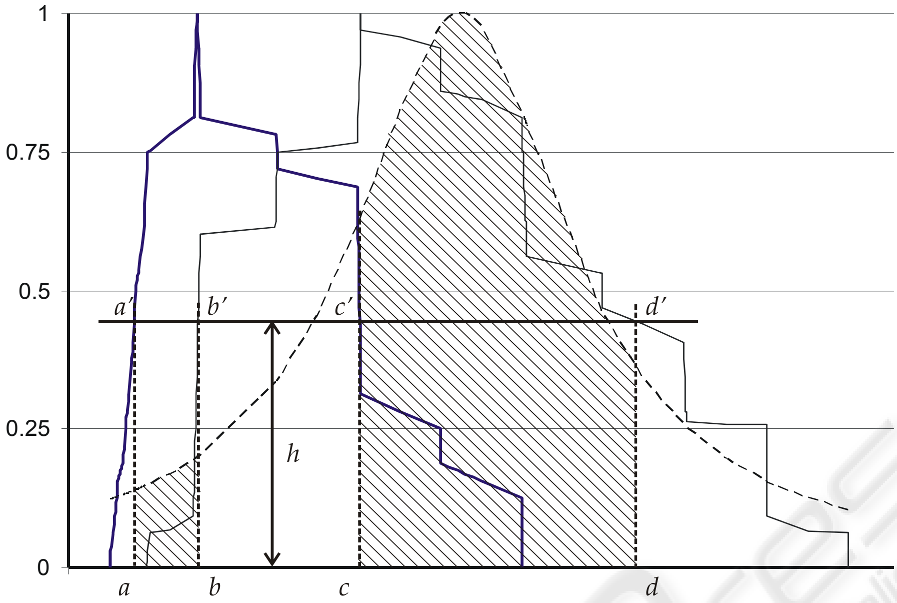

Figure 1: Two documents (vectors) illustrated as two Fuzzy Interval Numbers. Each value on the term axis represents a term (stem). Given two FINs any "cut" at a given height h∈ (0,1] defines two generalized intervals, denoted by [a', c']^h, [b', d']^h. In our case the generalized intervals are positive and intersecting. If you consider a "cut" at another height h=0.25 (∈ (0,1]) which defines a generalized interval denoted by F(h) or [b, d]$^{0.25}$ then the area [bd b' d'] is the support (F(0.25)) where there are about 75% of the values. Calculation of the distance between two FINs (or the similarity between two documents which are represented by their FINs). Use of a bell-shaped mass function for the calculations.

Figure 1 illustrates how we calculate the distance of two FINs F_1 and F_2 (representing two documents) using a mass function. The points *a*, *b*, *c*, *d* are used to define the distance, at height *h,* $d_h(F_1(h),F_2(h)) = d_h([a,b]^h,[c,d]^h)$; $d_h(F_1(h),F_2(h))$ equals the sum of areas of the shaded regions.

Hence, the distance of the two FINs at the height h is given by the sum of areas of the shaded regions, and eventually, the distance between the two FINs is calculated using the definite integral of the distance at height h from h=0 to 1:

$$\text{Distance} = \int_0^1 (|\int_{a'}^{b'} f(t)dt| + |\int_{c'}^{d'} f(t)dt|)dh$$

The FIN distance is used instead of the similarity measure between documents: the smaller the distance the more similar the documents. For the distance calculations a bell-shaped mass function was used:

$$\mathrm{m_h(t)}=\frac{\rho+(1-\rho)h}{1+\frac{1-z}{z}\left(\frac{t}{maxctf\,/\,v}-1\right)^2}[\sigma+(1-\sigma)t]$$

The positive real numbers ρ, σ, z, t, v are parameters; maxctf is the maximum value of all the collection term frequencies.

The mass function used was defined as tuning result of our experimentation with the "similar" bell-shaped mass function proposed by Marinagi et al. We used visual representations of mass-functions in order to study and improve our calculations. You can see some examples of our experimentation with such visual representations.

4.1 Illustrating Our Techniques by a First Simplified Experiment

The retrieved documents were extracted from a bibliographic database which is part of the Greek National archive of Dissertations. Each document (dissertation) is described by the title(s), the abstract(s), the key-phrases etc. All these fields of the description are bilingual: text in Greek and English (or other language e.g. French). Unfortunately, there are bibliographic descriptions that are not complete e.g. abstracts are not included in some cases. Our sample was constructed as the union of the retrieved documents by two simple queries (searches). More precisely, the documents of the National archive of Dissertations were searched using the (search) term "Natural Language Processing" and retrieved seven documents. Five documents contain the search term in their key-phrases and the other documents contain the term in other fields e.g. the abstract. Using the search term

Table 1: Vector space model for a collection of documents.

terms	ctf	#docs	Doc1	Doc2	Doc3	Doc4	Term identifiers
Term1	$ctf_1=3$	2	$tf_{1\,1}=0$	$tf_{2\,1}=1$	$tf_{3\,1}=0$	$tf_{4\,1}=2$	3
Term2	$ctf_2=3$	2	$tf_{1\,2}=0$	$tf_{2\,2}=2$	$tf_{3\,2}=1$	$tf_{4\,2}=0$	3,333
Term3	$ctf_3=3$	3	$tf_{1\,3}=0$	$tf_{2\,3}=1$	$tf_{3\,3}=1$	$tf_{4\,3}=1$	3,667
Term4	$ctf_4=4$	2	$tf_{1\,4}=0$	$tf_{2\,4}=1$	$tf_{3\,4}=0$	$tf_{4\,4}=3$	4
Term5	$ctf_5=4$	2	$tf_{1\,5}=3$	$tf_{2\,5}=0$	$tf_{3\,5}=1$	$tf_{4\,5}=0$	4,333
Term6	$ctf_6=4$	3	$tf_{1\,6}=2$	$tf_{2\,6}=1$	$tf_{3\,6}=0$	$tf_{4\,6}=1$	4,667
Term7	$ctf_7=5$	3	$tf_{1\,7}=1$	$tf_{2\,7}=3$	$tf_{3\,7}=0$	$tf_{4\,7}=1$	5
Term8	$ctf_8=5$	3	$tf_{1\,8}=3$	$tf_{2\,8}=1$	$tf_{3\,8}=0$	$tf_{4\,8}=1$	5.25
Term9	$ctf_9=5$	3	$tf_{1\,9}=2$	$tf_{2\,9}=0$	$tf_{3\,9}=2$	$tf_{4\,9}=1$	5,5
Term10	$ctf_{10}=5$	3	$tf_{1\,10}=1$	$tf_{2\,10}=2$	$tf_{3\,10}=2$	$tf_{4\,10}=0$	5,75
Term11	$ctf_{11}=6$	3	$tf_{1\,11}=1$	$tf_{2\,11}=0$	$tf_{3\,11}=2$	$tf_{4\,11}=3$	6
Term12	$ctf_{12}=7$	2	$tf_{1\,12}=0$	$tf_{2\,12}=6$	$tf_{3\,12}=0$	$tf_{4\,12}=1$	7
Term13	$ctf_{13}=8$	4	$tf_{1\,13}=2$	$tf_{2\,13}=2$	$tf_{3\,13}=1$	$tf_{4\,13}=3$	8
Term14	$ctf_{14}=8$	4	$tf_{1\,14}=3$	$tf_{2\,14}=2$	$tf_{3\,14}=1$	$tf_{4\,14}=2$	8,5
Term15	$ctf_{15}=9$	3	$tf_{1\,15}=1$	$tf_{2\,15}=4$	$tf_{3\,15}=0$	$tf_{4\,15}=4$	9

"Information Retrieval" we retrieved twelve documents and six of them contain the search term in their key-phrases. The FIN-based calculation of the distance (similarity) between the documents was used.

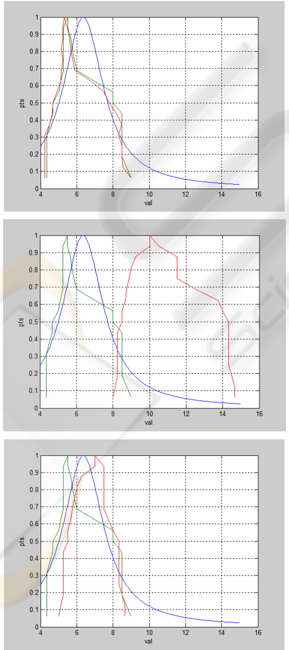

Figure 2: Tuning of the mass function using visualization.

Description of the Classification technique

The results of the two searches form a parallel corpus comprising two sets (classes) of documents (dissertations):

{8011, 8432, 8433, 8728, 11137, 11646, 12648}

{909, 2792, 3015, 3171, 7781, 8553, 8556, 11143, 12197, 12424, 13239, 13290}

Then, we can form two sub-collections:

NLP-Collection = {8011, 8432, 8433, 11137, 12648}

IR-Collection = {3171, 8553, 8556, 11143, 12424, 13290}

These collections include only the documents that contain the search term in their key-phrases. We know that in the first NLP-Collection other two documents are included: The document 8728 which has a detailed (bibliographic) description, and the document 11646 which has a shorter one (e.g. no abstract is included).

We measure the distance of these two documents from the two collections using the proposed FIN-based technique and examine if the documents are "closer" to the documents of the NLP-Collection. Using a different terminology we use the two collections as training sets and we try to classify correctly the documents 8728, 11646. Then, we must focus on the documents of the set {909, 2792, 3015, 7781, 12197, 13239} which are retrieved by the second query and measure the distance of these documents from the two collections using the proposed FIN-based technique, and test if they have to be classified in the NLP-collection.

We can calculate the average "matching" of the document 8728 from the NLP-collection: (0.105721+0.0670487+0.235943+0.213279+0.256549+0.240404)/6=0.1865

If we calculate the distance of the document 8728 from the IR-Collection = {3171, 8553, 8556, 11143, 12424, 13290} or the broader collection {909, 2792, 3015, 3171, 7781, 8553, 8556, 11143, 12197, 12424, 13239, 13290} (and we can also use or not stop-words file) we conclude that in general it looks reasonable to add (classify) the document 8728 in the NLP-Collection. The details of the experiment are summarized in the Table II below. As you can see the document 11646 is correctly classified in the NLP-Collection. We conclude that the validity of the FIN-based calculation of the similarity between documents is verified in the case of the documents (dissertations) that are related to the search term "natural language processing".

Table 2: Experimental details.

Document	IR-collection	NLP-Collection	Comments
8728	0.3228	0.1865	Correct
11646	0.6788	0.2606	Correct

When the documents were restricted only in the English part (e.g. title and abstract only in English) the documents' length was reduced and some documents were erroneously classified. This was an indication that our technique is more appropriate in the case of "large" documents. In some cases there was a need to "correct" the classification errors (and in general improve our results) using a technique of weights (or penalties or comparison with the top-x "hits") as we can see in Table III. Some improvement could be also mentioned in the case of adding in the collections "artificial" documents (centroids) comprising elements of the documents. We had also indications that such techniques will be more useful in the case of bigger documents. The same experiment was also conducted and all the documents without abstracts were removed. Only the documents {8011, 8432, 8433, 8728}, {909, 2792, 3015, 3171, 7781, 8553, 8556, 12197} were used and we can report some improvement. Then we "split" every document in two documents: the Greek part of the document and the English one e.g. 8728Gre, 8728Eng. In table III we can see how we can improve our classification results.

Table 3: Improved classification results.

Document number	Distance from IR-collection	Distance from NLP-Collection	Comments
8728Gre	0.5889	0.1507	
8728Eng	0.7551	0.1503	
3171Gre	0.1088	0.1451	use of weights
3171Eng	0.2904	0.2922	use of weights
8553Gre	0.1773	0.2220	use of weights
8553Eng	0.2141	0.3833	
8553Eng	<=0.0680	0.0776	use of the top 4/5 "hits" only

Table 4: Brief experimental corpora description.

Title and Author	Lang
Common Sense Parenting (CSP) Learn at home kit: A clinical effectiveness evaluation of a commercially available video training program for parents. Sean T. Smitham, Ph.D., Western Michigan University, 2004	English
The effects of group size on incentive effectiveness: A meta-analysis. Angelica C. Grindle, Ph.D., Western Michigan University, 2002	English
A Market Analysis of consumer behaviour for companies in a self-insurance group, Bismarck J. Manes Jr., M.A., Western Michigan University, 2006	English
Refinement of temporal constraints in an event recognition system using small datasets, George Paliouras (NCSR "Demokritos", Greece)	English
A Formal Semantics for the C Programming Language, Nikolaos S. Papaspyrou (National Technical University of Athens)	English
Updating and retrieving information through relational database views, Panayiota Plessa (Ministry of Education, Greece)	Greek
The role of asynchronous hypermedia conferencing in education and training, Cleo Sgouropoulou (TEI of Athens)	Greek
Information system's Security management for coalitions in distributed environments Petros Belsis (TEI of Athens)	Greek

5 CONCLUSIONS

We have concluded that the FIN-based calculation of the similarity between documents is a novel method for solving various problems in the case of CLIR-systems. We verified in our experiments that if we focus on the documents (dissertations etc) that are related to different search terms then we can apply FIN-based techniques and calculate correctly the distance (similarity) between documents.

Such techniques include the following cases:

1. Use partitions ("collections") of the sample (e.g. you can use the NLP-Collection and the IR-Collection) and calculate the distance of the ("unclassified") document from such partitions ("collections"). This distance is some kind of average distance of the document from all the elements of the collection.
2. Use partitions ("collections") of the sample, and define the number of "hits" (e.g. top-4 or top-5). Calculate the distances of the ("unclassified") document from the (e.g. four or five) documents of each partition. Use only the (top-x) documents which are closer to the unclassified one. Then you can calculate the average distance of the unclassified document from these top-x documents.
3. Use increased weights for the search terms that are contained in the retrieved documents.
4. Use of positive weights in the case that the search term is included in a document of the partition and use of penalty (negative weight) in the case that the search term is not included.
5. If the length of the documents is greater then the results of the method are better.

A strategy related to the specific sample of classified and unclassified documents could be defined. As an example, you can combine 1 & 2 and if it is necessary weights and penalty: If the general (average) distance of a document from a collection (which is calculated from the distances of the document from all the classified documents of the collection) and the partial one (which is calculated from a number of the top-x distances) are not "consistent" you must use weights following the appropriate technique.

ACKNOWLEDGEMENTS

This work was co-funded by 75% from the European Social Fund and 25% by National Resources (EPEAEK-II)-Archimedes.

REFERENCES

Radecki,T (1979), "Fuzzy Set Theoretical Approach to Document Retrieval" in Information Processing and Management, v.15, Pergammon-Press 1979.

Kraft, D.H. and D.A. Buell (1993), "Fuzzy Set and Generalized Boolean Retrieval Systems" in Readings in Fuzzy Sets for Intelligent Systems, D. Dubius, H.Prade, R.R. Yager (eds).

Kaburlasos, V.G. (2004), "Fuzzy Interval Numbers (FINs): Lattice Theoretic Tools for Improving Prediction of Sugar Production from Populations of Measurements," IEEE Trans. on Man, Machine and Cybernetics – Part B, vol. 34, no 2, pp. 1017-1030.

Petridis, V. and V.G. Kaburlasos (2003), "FINkNN: A Fuzzy Interval Number k-Near-est Neighbor Classifier for prediction of sugar production from populations of samples," Journal of Machine Learning Research, vol. 4 (Apr), pp. 17-37, 2003

Kaburlasos and Petridis (2000), Fuzzy Lattice Neurocomputing models, Neural Networks, 13(10), 1145-1170.

Petridis and Kaburlasos (1998), Fuzzy lattice neural network (FLNN): A hybrid model for learning, IEEE Trans. Neural Networks, 9(5), 877-890.

Petridis and Kaburlasos (2000), An intelligent mechatronics solution for automated tool guidance in the epidural surgical procedure, Proc. 7th Annual conf. Mechatronics and Machine Vision in Practice, pp 201-206.

Petridis and Kaburlasos (2001), Clustering and classification in structured data domains using Fuzzy Lattice Neurocomputing, IEEE Trans. Knowledge Data Engineering, 13(2), 245-260, 2001

Kaburlasos et al (1997), Automatic detection of bine breakthrough in orthopedics by fuzzy lattice reasoning: The case of drilling in the osteosynthesis of long bones, Proc. Mechatronics Computer systems for Perception and Action, pp 33-40.

Athanassiadis and Mitkas (2003), Applying machine learning techniques on air quality data for real-time decision support, Proc. Intl. NAISO Symposium on Information Technologies in Environmental Engineering.

Kaburlasos V.G., Spais V, Petridis V, Petrou L, Kazarlis S, Maslaris N, and Kallinakis A, Intelligent clustering techniques for prediction of sugar production, Mathematics and Computers in Simulation, 60(3-5), 159-168, 2002

Kaburlasos V.G. Papadakis S. (2005) granular Self Organizing Map (grSOM) neural network for industrial quality control, Proc of SPIE, Mathematical Methods in Pattern and Image Analysis, 2005

Kaburlasos V.G. , Fuzzy Interval Numbers (FINs): Lattice Theoretic Tools for Improving Prediction of Sugar Production from Populations of Measurements

Marinagi, Alevisos, Kaburlasos, Skourlas, Fuzzy Interval Number (FIN) Techniques for Cross Language Information Retrieval, Proc. 8th ICEIS, 2006