Managing Unbalanced Linguistic Information in an Information Retrieval System[0]

E. Herrera-Viedma, A.G. López-Herrera, C. Porcel
Dept. of Computer Science and A.I.
University of Granada,
18071 - Granada
viedma@decsai.ugr.es

L. Villen
Library Sciences School
University of Granada,
18071 - Granada

Abstract

Most information retrieval systems based on linguistic approaches use symmetrically and uniformly distributed linguistic term sets to express the weights of queries and the relevance degrees of documents. However, it seems more adequate to express these linguistic weights and degrees by means of unbalanced linguistic scales, i.e., linguistic term sets with different discrimination levels on both sides of mid linguistic term. In this contribution we present an information retrieval system which accepts weighted queries whose weights are expressed using unbalanced linguistic term sets. Then, system provides the retrieved documents classified in linguistic relevance classes assessed on unbalanced linguistic term sets. To do so, we introduce a methodology to manage unbalanced linguistic information which is composed of a representation model of unbalanced linguistic information and a computational model of unbalanced linguistic information with some aggregation operators.

Keywords: Information Retrieval. Weighted Query. Unbalanced Linguistic Information.

[0]This research has been supported by CICYT under Project TIC2003-07977.

1 Introduction

Information Retrieval involves the development of computer systems for the storage and retrieval of (predominantly) textual information (documents). The main activity of an Information Retrieval System (IRS) is the gathering of the pertinent filed documents that best satisfy user information requirements (queries). Basically, IRSs present three components to carry out their activity [21]: i) a *Database:* which stores the documents and the representation of their information contents (index terms); ii) a *Query Subsystem:* which allows users to formulate their queries by means of a query language; and iii) an *Evaluation Subsystem:* which evaluates the relevance of each document for a user query by means of a retrieval status value (RSV).

A promising direction to improve the effectiveness of IRSs consists of representing in the queries the users' concept of relevance. To do so, a possible solution consists in the use of weighting tools in the formulation of queries. By attaching weights in a query, a user can increase his/her expressiveness and provide a more precise description of his/her desired documents.

The *fuzzy linguistic approach* is an approximate tool to model qualitative information in problems [1, 2, 14, 15, 18, 19]. An important fuzzy linguistic approach is called the *ordinal fuzzy linguistic approach* [7, 9, 24]. Its main characteristic is that it simplifies the processes of computing with words [7]. So different weighted IRSs based on an ordinal

fuzzy linguistic approach were presented in [3, 4, 12, 13, 16, 17]. With such linguistic approach the weights are assumed qualitative values assessed on symmetrically and uniformly distributed linguistic term sets. Then, users can characterize the contents of the desired documents by explicitly associating a linguistic descriptor to a term in a query, such as "important" or "very important", and on the other hand, the estimated relevance levels of the documents are supplied in a linguistic form (e.g., linguistic terms such as "relevant", "very relevant" may be used). The problem is that using symmetrically and uniformly distributed linguistic term sets we find the same discrimination levels on both sides of mid linguistic term. However, usually users look for documents with positive criteria, that is, they formulate their weighted queries using linguistic assessments on the right of the mid label a lot more than on the left. Similarly, usually users are interested in the relevant documents a lot more than in the non-relevant documents, and then a best tuning of the output of IRS can be achieved if a higher number of discrimination levels on the right of the mid linguistic term is assumed. Therefore, in information retrieval the use of *unbalanced linguistic term sets* (see Figure 1) i.e., linguistic term sets with different discrimination levels on both sides of the mid linguistic term, seems more appropriate to express weighted queries and the relevance of documents.

Figure 1: Example of an unbalanced set of 7 fuzzy linguistic terms.

The aim of this contribution is to present a linguistic IRS that manages unbalanced fuzzy linguistic information to represent the weights of queries and the relevance degrees of retrieved documents. To do so, we introduce a new methodology to manage the unbalanced fuzzy linguistic information with both its representation model of linguistic information and its computational model to aggregate unbalanced fuzzy linguistic information. This methodology is based on the use of hierarchical linguistic contexts [5] defined using a 2-tuple fuzzy linguistic approach[10]. In such a way, we present an IRS that improves the expressiveness in the system-user interaction. Furthermore, the use of 2-tuple model improves the performance of IRS because it allows to represent more classification levels of relevance in the output of IRS.

In order to do that, this contribution is structured as follows. Section 2 shows the preliminary concepts, that is, the 2-tuple fuzzy linguistic approach and hierarchical linguistic contexts. Section 3 introduces the methodology designed to manage unbalanced fuzzy linguistic information. Section 4 defines the linguistic IRS based on unbalanced fuzzy linguistic information. And finally, some concluding remarks are pointed out.

2 Preliminaries

In this section, we present those concepts that we use to design the methodology to manage unbalanced fuzzy linguistic information and our linguistic IRS. That is, we present the 2-tuple fuzzy linguistic model [10] and the hierarchical linguistic contexts [5].

2.1 The 2-Tuple Fuzzy Linguistic Model

Usually, to define a fuzzy linguistic model we must establish its representation model of linguistic information and its computational model to combine linguistic information. In such a way, in [10] was introduced the 2-tuple fuzzy linguistic model by defining both its representation model and its computational model.

Definition 1. *Let $S = \{s_0, ..., s_T\}$ be a linguistic term set and $\beta \in [0, T]$ a value supporting the result of a symbolic aggregation operation, then the 2-tuple that expresses the equivalent information to β is obtained with the following function:*

$$\Delta : [0, T] \longrightarrow S \times [-0.5, 0.5)$$

$$\Delta(\beta) = (s_i, \alpha), \begin{cases} s_i & i = round(\beta) \\ \alpha = \beta - i & \alpha \in [-.5, .5) \end{cases}$$

where round(·) is the usual round *operation, $s_i \in S$ has the closest index label to "β" and "α" is the value of the symbolic translation.*

Proposition 1. *Let $S = \{s_0, ..., s_T\}$ be a linguistic term set. There is always a Δ^{-1} function, such that, from a 2-tuple (s_i, α) it returns its equivalent numerical value $\beta \in [0, T] \subset \mathcal{R}$.*

$$\Delta^{-1} : S \times [-.5, .5) \longrightarrow [0, T]$$

$$\Delta^{-1}(s_i, \alpha) = i + \alpha = \beta$$

Remark 1. We should point out that the conversion of a linguistic term into a linguistic 2-tuple consists of adding a value 0 as value of symbolic translation: $s_i \in S \Longrightarrow (s_i, 0)$.

On the other hand, the 2-tuple fuzzy linguistic computational model carries out processes of computing with words in a precise way when the linguistic term sets are symmetrically and uniformly distributed. This computational model presents different techniques to manage the linguistic information [10]:

- *Comparison of 2-tuples:* The comparison of linguistic information represented by 2-tuples is carried out according to an ordinary lexicographic order. Let (s_k, α_1) and (s_l, α_2) be two 2-tuple:
  - if $k < l$ then (s_k, α_1) is smaller than (s_l, α_2)
  - if $k = l$ then
    - if $\alpha_1 = \alpha_2$ then (s_k, α_1), (s_l, α_2) represent the same information
    - if $\alpha_1 < \alpha_2$ then (s_k, α_1) is smaller than (s_l, α_2)
    - if $\alpha_1 > \alpha_2$ then (s_k, α_1) is bigger than (s_l, α_2)
- *Negation of 2-tuple:* is defined as

$$Neg(s_i, \alpha) = \Delta(T - \Delta^{-1}(s_i, \alpha)).$$

- *Aggregation of 2-tuples:* Using the function Δ and Δ^{-1} any aggregation operator can be easily extended for dealing with linguistic 2-tuples. Some examples are presented in [10].

2.2 Hierarchical Linguistic Contexts Based on Fuzzy Linguistic 2-Tuples

The hierarchical linguistic contexts were introduced in [5] to improve the linguistic modelling in fuzzy systems. They were used in [11] to improve the precision of processes of computation with words in multi-granular fuzzy linguistic contexts. In this contribution, we use them to manage unbalanced fuzzy linguistic term sets.

A *Linguistic Hierarchy* is a set of levels, where each level represents a linguistic term set with different granularity to the remaining levels. Each level is denoted as $l(t, n(t))$, being, t a number that indicates the level of the hierarchy, and $n(t)$ the granularity of the linguistic term set of the level t.

We assume levels containing linguistic terms whose membership functions are triangular-shaped, symmetrically and uniformly distributed in $[0, 1]$. In addition, the linguistic term sets have an odd value of granularity.

The levels belonging to a linguistic hierarchy are ordered according to their granularity, i.e., for two consecutive levels t and $t+1$, $n(t+1) > n(t)$. Therefore, the level $t+1$ is a refinement of the previous level t.

From the above concepts, we define a linguistic hierarchy, $\mathcal{LH}$, as the union of all levels t: $\mathcal{LH} = \bigcup_t l(t, n(t))$.

Given an $\mathcal{LH}$, we denote as $S^{n(t)}$ the linguistic term set of $\mathcal{LH}$ corresponding to the level t of $\mathcal{LH}$ characterized by a granularity of uncertainty $n(t)$: $S^{n(t)} = \{s_0^{n(t)}, ..., s_{n(t)-1}^{n(t)}\}$.

Generically, we can say that the linguistic term set of level $t+1$ is obtained from its predecessor as: $l(t, n(t)) \rightarrow l(t+1, 2 \cdot n(t) - 1)$.

In [11] transformation functions between labels of different levels were developed to make processes of computing with words without loss of information.

Definition 2. *Let $\mathcal{LH} = \bigcup_t l(t, n(t))$ be a linguistic hierarchy whose linguistic term sets are denoted as $S^{n(t)} = \{s_0^{n(t)}, ..., s_{n(t)-1}^{n(t)}\}$, and*

let us consider the 2-tuple linguistic representation. The transformation function from a linguistic label in level t to a label in level t' is defined as: $TF_{t'}^{t} : l(t, n(t)) \longrightarrow l(t', n(t'))$

$$TF_{t'}^{t}(s_i^{n(t)}, \alpha^{n(t)}) =$$

$$\Delta_{n(t')}\left(\frac{\Delta_{n(t)}^{-1}(s_i^{n(t)}, \alpha^{n(t)}) \cdot (n(t') - 1)}{n(t) - 1}\right).$$

Proposition 2. The transformation function between linguistic terms in different levels of the linguistic hierarchy is bijective:

$$TF_{t}^{t'}(TF_{t'}^{t}(s_i^{n(t)}, \alpha^{n(t)})) = (s_i^{n(t)}, \alpha^{n(t)}).$$

3 A Methodology to Manage the Unbalanced Fuzzy Linguistic Information

Here, we propose a method to manage unbalanced linguistic term sets based on the linguistic 2-tuple model. Basically, this method consists of representing unbalanced linguistic terms from different levels of a $\mathcal{LH}$, carrying out computational operations of unbalanced linguistic information using the 2-tuple computational model.

The management method of unbalanced linguistic information presents the following steps:

3.1 Representation the unbalanced linguistic term set $\mathcal{S}$ by means of a linguistic hierarchy $\mathcal{LH}$

To do this, we use different levels of the linguistic hierarchy $\mathcal{LH}$ to represent both sides of the mid linguistic term. So, the side with more linguistic terms needs a more granular level $l(i, n(i))$ of $\mathcal{LH}$ and the side with less linguistic terms needs a lees granular level $l(j, n(j))$ of $\mathcal{LH}$, being $i > j$. Concretely, the steps are:

1.- Choose a level t^- with an adequate granularity to represent using the 2-tuple representation model the subset of linguistic terms of $\mathcal{S}$ on the left of the mid linguistic term, and

2.- Choose a level t^+ with an adequate granularity to represent using the 2-tuple representation model the subset of linguistic terms of $\mathcal{S}$ on the right of the mid linguistic term.

3.2 Define an unbalanced linguistic computational model

To manage unbalanced linguistic information we need a computation tools set, so, in the following points we describe some basic tools:

1.- Choose a level $t' \in \{t^-, t^+\}$, such that $n(t') = max\{n(t^-), n(t^+)\}$.

2.- Define the comparison of two unbalanced 2-tuples $(s_k^{n(t)}, \alpha_1)$, $t \in \{t^-, t^+\}$, and $(s_l^{n(t)}, \alpha_2)$, $t \in \{t^-, t^+\}$. Its expression is similar to the usual comparison of two 2-tuples but acting on the values $TF_{t'}^{t}(s_k^{n(t)}, \alpha_1)$ and $TF_{t'}^{t}(s_l^{n(t)}, \alpha_2)$. We should point out that using the comparison of unbalanced 2-tuples we can easily define the comparison operators Max_{un} and Min_{un}.

3.- Define the negation operator of unbalanced linguistic information. Let $(s_k^{n(t)}, \alpha)$, $t \in \{t^-, t^+\}$ be an unbalanced 2-tuple then:

$$\mathcal{NEG}(s_k^{n(t)}, \alpha) = Neg(TF_{t''}^{t}(s_k^{n(t)}, \alpha)),$$

$t \neq t'', t'' \in \{t^-, t^+\}$.

4.- Define aggregation operators of unbalanced linguistic information. This is done using the aggregation processes designed in the 2-tuple computational model but acting on the unbalanced linguistic values transformed by means of $TF_{t'}^{t}$. Then, once it is obtained a result, it is transformed to the correspondent level $t \in \{t^-, t^+\}$ by means of $TF_{t}^{t'}$ for expressing the result in the unbalanced linguistic term set.

For example, we can easily define the $LOWA_{un}$ operator, which is an extension of the $LOWA$ defined in [9] as follows:

Definition 3. *Let $\{(a_1, \alpha_1), \ldots, (a_m, \alpha_m)\}$ be a set of unbalanced assessments to aggregate, then the $LOWA_{un}$ operator ϕ_{un} is defined as:*

$$\phi_{un}((a_1, \alpha_1), \ldots, (a_m, \alpha_m)) = W \cdot B^T =$$

$$= C_{un}^{m}\{w_w, b_w, k = 1, \ldots, m\} =$$

$w_1 \otimes b_1 \oplus (1-w_1) \otimes C_{un}^{m-1}\{\beta_h, b_h, h = 2, \ldots, m\}$ *where* $b_i = (a_i, \alpha_i) \in (S \times [-.5, .5)), W = [w_1, \ldots, w_m]$, *is a weighting vector, such that,* $w_i \in [0,1]$ *and* $\sum_i w_i = 1$, $\beta_h = \frac{w_h}{\sum_2^m w_k}, h = 2, \ldots, m$, *and* B *is the associated ordered unbalanced 2-tuple vector. Each element* $b_i \in B$ *is the i-th largest unbalanced 2-tuple in the collection* $\{(a_1, \alpha_1), \ldots, (a_m, \alpha_m)\}$, *and* C_{un}^m *is the convex combination operator of* m *unbalanced 2-tuples. If* $w_j = 1$ *and* $w_i = 0$ *with* $i \neq j \forall i, j$ *the convex combination is defined as:* $C_{un}^m\{w_i, b_i, i = 1, \ldots, m\} = b_j$. *And if* $m = 2$ *then it is defined as:* $C_{un}^2\{w_l, b_l, l = 1, 2\} = w_1 \otimes b_j \oplus (1-w_1) \otimes b_i = TF_t^{t'}(s_k^{n(t')}, \alpha)$, *where* $(s_k^{n(t')}, \alpha) = \Delta(\lambda)$ *and* $\lambda = \Delta^{-1}(TF_{t'}^t(b_i)) + w_1 \cdot (\Delta^{-1}(TF_{t'}^t(b_j)) - \Delta^{-1}(TF_{t'}^t(b_i)))$, $b_j, b_i \in (S \times [-.5, .5))$, $(b_j \geq b_i)$, $\lambda \in [0, n(t')-1], t \in \{t^-, t^+\}$.

We also can define a weighted operator to aggregate weighted unbalanced linguistic information.

Usually, a weighted aggregation operator to aggregate information carries out two activities [8]:

1.- Transformation of the weighted information under the weighted degrees by means of a transformation function h. Examples of families of connectives used as transformation functions are the following two:

a.) *Linguistic conjunction functions* ($LC^{\rightarrow}$). The linguistic conjunction functions that we shall use are the following t-norms, which are monotonically nondecreasing in the weights and satisfy the properties required for any transformation function, h, [6]: i) the classical *MIN* operator, ii) the *nilpotent MIN* operator, and iii) the *weakest conjunction*.

b.) *Linguistic implication functions* ($LI^{\rightarrow}$). The linguistic implication functions that we shall use are monotonically nonincreasing in the weights and satisfy the properties required for any transformation function h [6]: i) *Kleene-Dienes's* implication function, ii) *Gödel's* implication function, and iii) *Fodor's* implication function.

2.- The aggregation of the transformed weighted information by means of an aggregation operator of non-weighted information f. As it is known, the choice of h depends upon f. As f operator we can use the $LOWA_{un}$ with the transformed weighted degrees by h.

In order to classify OWA operators ($LOWA_{un}$ operator is based in OWA) in regards to their location between "and" and "or" Yager [23] introduced an *orness measure*[1] associated with any vector W, which allows to control its aggregation behavior.

4 The IRS with Unbalanced Linguistic Information

In this section we present a linguistic IRS which uses an unbalanced linguistic term set S to express the linguistic assessments in the retrieval process. Particularly, S presents a higher number of discrimination levels on the right of the mid linguistic term than on the left (as happens in example of Figure 1). Then, this IRS accepts linguistic weighted queries and provides linguistic retrieval status values (RSVs) assessed on S and $S \times [-.5, .5)$, respectively. The components of this IRS are presented in the following subsections.

4.1 Database

The database stores the finite set of documents $\mathcal{D} = \{d_1, \ldots, d_m\}$ and the finite set of index terms $\mathcal{T} = \{t_1, \ldots, t_l\}$. Documents are represented by means of index terms, which describe the subject content of the documents. A numeric indexing function $\mathcal{F}$: $\mathcal{D} \times \mathcal{T} \rightarrow [0,1]$, exists. $\mathcal{F}$ weighs index terms according to their significance in describing the content of a document in order to improve the retrieval of documents. We assume that the system uses any of the existing weighting methods [21] to compute $\mathcal{F}$.

4.2 The Query Subsystem

The query subsystem presents a weighted Boolean query language to express user information needs. In the queries, the terms can

[1] $orness(W) = \frac{1}{m-1} \sum_{i=1}^m (m-i) \cdot w_i$.

be weighted according to two different semantics possibilities, even simultaneously. These semantics are a threshold semantics and a relative importance semantics. As in [2] we use the linguistic variable *Importance* to express the linguistic weights associated to the query terms. Thus, we consider a set of unbalanced linguistic values $\mathcal{S}$.

By associating threshold weights with terms in a query, the user is asking to see all the documents sufficiently about the topics represented by such terms. By associating importance weights to terms in a query, the user is asking to see all documents whose content represents the concept that is more associated with the most important terms than with the less important ones. Each query is expressed as a combination of the weighted index terms which are connected by the logical operators AND ($\wedge$) and OR ($\vee$).

Therefore, a query $\mathcal{Q}$ is any legitimate Boolean expression whose atomic components (atoms) are 3-tuples $< t_i, c_i^1, c_i^2 >$ belonging to the set, $\mathcal{T} \times \mathcal{S}^2$; $t_i \in \mathcal{T}$, and c_i^1 and c_i^2 are linguistic values of the linguistic variable *Importance* modelling the threshold semantics (importance that the term t_i must have in the desired documents) and importance semantics (importance that the meaning of t_i must have in the set of retrieved documents), respectively. Accordingly, the set $\mathcal{Q}$ of the legitimate queries is defined by the following syntactic rules:

1.- $\forall q =< t_i, c_i^1, c_i^2 > \in \mathcal{T} \times \mathcal{S}^2 \longrightarrow q \in \mathcal{Q}$.

2.- $\forall q, p \in \mathcal{Q} \longrightarrow q \wedge p, q \vee p \in \mathcal{Q}$.

3.-All legitimate queries $p \in \mathcal{Q}$ are only those obtained by applying rules 1-2, inclusive.

4.3 The Evaluation Subsystem

The goal of an *evaluation subsystem* consists of evaluating documents in terms of their relevance to a linguistic weighted Boolean query according to two possible semantics. A Boolean query with more than one weighted term is evaluated by means of a constructive bottom-up process which includes the following four steps:

1.- *Preprocessing of the query:* In this step, the user query is preprocessed and put into either *conjunctive normal form* (CNF) or *disjunctive normal form* (DNF), with the result that all its Boolean subexpressions must have more than two atoms.

2.- *Evaluation of atoms with respect to the threshold semantics:* In this step, the documents are evaluated with regard to their relevance to individual atoms in the query, considering only the restrictions imposed by the threshold semantics. To model the interpretation of the threshold semantics, we use the matching function described in [20] but defined in a 2-tuple unbalanced linguistic context, it is called g_{un}, and defined as: $g_{un} : \mathcal{D} \times \mathcal{T} \times \mathcal{S} \longrightarrow \mathcal{S} \times [-.5, .5)$. Then, given an atom $< t_i, c_i^1, c_i^2 >$, $t_i \in \mathcal{T}$, and $d_j \in \mathcal{D}$, g_{un} obtains the partial 2-tuple linguistic RSV of d_j, called $RSV_j^{i,1}$, by measuring how well the index term weight $\mathcal{F}(d_j, t_i)$ satisfies the request expressed by the linguistic threshold weight c_i^1 according to the following expression:

$$g_{un}(d_j, t_i, c_i^1) = \begin{cases} (s_a, \alpha_a) & \text{if } (s_a, \alpha_a) \geq (c_i^1, 0) \\ \Delta(0) & \textit{otherwise.} \end{cases}$$

where $(s_a, \alpha_a) = \Delta((n(t) - 1) \cdot \mathcal{F}(d_j, t_i))$, $\Delta : [0, n(t) - 1] \longrightarrow \mathcal{S} \times [-.5, .5)$ with $t = t^-$ if $\mathcal{F}(d_j, t_i) \leq .5$ and $t = t^+$ if $\mathcal{F}(d_j, t_i) > .5$, being t^- and t^+ the levels of $\mathcal{LH}$.

3.- *Evaluation of subexpressions and modelling the importance semantics:* We consider that the relative importance semantics in a single-term query has no meaning. Then, in this step we have to evaluate the relevance of documents with respect to all subexpressions of preprocessed queries which are composed of a minimum number of two atomic components.

Given a subexpression q_v, with $\eta \geq 2$ atoms, we know that each document d_j presents a partial $RSV_j^{i,1} \in (\mathcal{S} \times [-.5, .5))$ with respect to each atom $< t_i, c_i^1, c_i^2 >$ of q_v. Then, the evaluation of the relevance of a document d_j with respect to the whole expression q_v implies the aggregation of the partial relevance degrees $\{RSV_j^{i,1}, i = 1 \ldots, \eta\}$ weighted by

means of the respective relative importance degrees $\{c_i^2 \in \mathcal{S}, i = 1, \ldots, \eta\}$. To do that, we need a weighted aggregation operator of 2-tuple linguistic information which should guarantee that the more important the query terms, the more important they are in the determination of the RSVs.

In [22], Yager discussed the effect of the importance degrees on the MAX and MIN types of aggregation and suggested a class of functions for importance transformation in both types of aggregation. For the MIN aggregation, he suggested a family of t-conorms acting on the weighted information and the negation of the importance degrees, for the MAX aggregation, he suggested a family of t-norms acting on weighted information and the importance degree. As it is known, the evaluation of the logical connectives AND and OR by means of the MIN and MAX operators presents some limitations. That is, it may cause a very restrictive and inclusive behavior, respectively. This fact provokes that the retrieval process may be deceptive because, on the one hand, the linguistic MIN t-norm may cause the rejection of useful documents by the dissatisfaction of any one single criterion of the conjunctive subexpression and, on the other hand, the linguistic MAX t-conorm may cause the acceptance of a useless document by the satisfaction of any single criterion.

Therefore, to aggregate weighted unbalanced linguistic information we use the unbalanced $LOWA_{un}$ operator ϕ_{un} together with the transformation functions $LC^{\rightarrow}$, the classical *MIN* operator, and $LI^{\rightarrow}$, *Gödel's* implication function, to model the weighted AND and OR Boolean connectives respectively. Futhermore, these operators overcome the above limitations of the linguistic t-norm MIN and t-conorm MAX because its behavior can be soften by means of the weighting vector.

Then, we use the orness measure to control the behavior of the $LOWA_{un}$ operator ϕ_{un}. In particular, we propose to use an unbalanced operator ϕ_{un}^1 with $orness(W) \geq .5$ to model the AND connectives and an unbalanced operator ϕ_{un}^2 with $orness(W) < .5$ to model the OR connective.

Hence, to evaluate the subexpressions together with the relative importance semantics and according to activities necessary to aggregate weighted information, if the subexpression is conjunctive then we use $f = \phi_{un}^1$ and $h = MAX_{un}(\mathcal{NEG}(weight, 0)$, unbalanced value), and if it is disjunctive then we use $f = \phi_{un}^2$, then $h = MIN_{un}((weight, 0)$, unbalanced value).

Shortly, given a document d_j, we evaluate its relevance with respect to a subexpression q_v, called $RSV_j^v \in (\mathcal{S} \times [-.5, .5))$ as:

1.- If q_v is a conjunctive subexpression then

$$RSV_j^v = \phi_{un}^1(MAX_{un}(\mathcal{NEG}(c_1^2, 0), RSV_j^{1,1}), \ldots, MAX_{un}(\mathcal{NEG}(c_\eta^2, 0), RSV_j^{\eta,1})).$$

2.- If q_v is a disjunctive subexpression then

$$RSV_j^v = \phi_{un}^2(MIN_{un}((c_1^2, 0), RSV_j^{1,1}), \ldots, MIN_{un}((c_\eta^2, 0), RSV_j^{\eta,1})).$$

4.- *Evaluation of the whole query:* In this final step of evaluation, the documents are evaluated with regards to their relevance to Boolean combinations in all the Boolean subexpressions existing in a query. To do that, we use again both unbalanced $LOWA_{un}$ operators ϕ_{un}^1 and ϕ_{un}^2 to model the AND and OR connectives, respectively.

Then, given a document d_j, its relevance with respect to a query q, $RSV_j \in (\mathcal{S} \times [-.5, .5))$ as:

1.- If q is in CNF then $RSV_v = \phi_{un}^1(RSV_j^1, \ldots, RSV_j^v)$

2.- If q is in DNF then $RSV_v = \phi_{un}^2(RSV_j^1, \ldots, RSV_j^v)$,

with v standing for the number of subexpressions of q.

Shortly, this evaluation subsystem can be synthesized by means of a general linguistic evaluation function $\mathcal{E}_{un} : \mathcal{D} \times \mathcal{Q} \longrightarrow \mathcal{S} \times [-.5, .5))$, which evaluates the different kind of preprocessed queries, $\{q =< t_i, c_i^1, c_i^2 >, q \wedge p, q \vee p\}$ according to the following five rules:

1.- *Atoms:* $\mathcal{E}$
that, $q^1 =< t$

2.- *Conjuncti*

$\phi_{un}^1(MA$

$\ldots, MAX$

being η the n

3.- *Disjunctiv*

$\phi_{un}^2(M$

$\ldots, M$

4.- *Query in*

$\phi_{un}^1(\mathcal{E}_i$

being ω the
sions.

5.- *Query in*

$\phi_{un}^1(\mathcal{E}_i$

Then, the iss
is a fuzzy su
by the lingui

$\{(d_1, \mathcal{E}_{un}($

The docume
der of $\mathcal{E}_{un}$ ar
classes, in su
ber of classe
the unbalan
sented the li

5 Conclu

In this cont
guistic IRS
sets. In such
use a higher
to assess the
of queries, a
has also a
values to as
retrieved dc
veloped a m

connective.

valuate the subexpressi

1.- *Atoms*: $\mathcal{E}_{un}(d_j, q^1) = g_{un}(d_j, t_i, c_i^1)$, such that, $q^1 =< t_i, c_i^1, c_i^2 >$.

2.- *Conjunctive subexpressions*: $\mathcal{E}_{un}(d_j, q^2) =$

$$\phi_{un}^1(MAX_{un}(\mathcal{NEG}(c_1^2, 0), \mathcal{E}_{un}(d_j, q_1^1)),$$

$$\ldots, MAX_{un}(\mathcal{NEG}(c_\eta^2, 0), \mathcal{E}_{un}(d_j, q_\eta^1))).$$

being η the number of atoms of q^2.

3.- *Disjunctive subexpressions*: $\mathcal{E}_{un}(d_j, q^3) =$

$$\phi_{un}^2(MIN_{un}((c_1^2, 0), \mathcal{E}_{un}(d_j, q_1^1)),$$

$$\ldots, MIN_{un}((c_\eta^2, 0), \mathcal{E}_{un}(d_j, q_\eta^1))),$$

4.- *Query in CNF*: $\mathcal{E}_{un}(d_j, q^4) =$

$$\phi_{un}^1(\mathcal{E}_{un}(d_j, q_1^3), \ldots, \mathcal{E}_{un}(d_j, q_\omega^3)),$$

being ω the number of disjunctive subexpressions.

5.- *Query in DNF*: $\mathcal{E}_{un}(d_j, q^5) =$

$$\phi_{un}^1(\mathcal{E}_{un}(d_j, q_1^2), \ldots, \mathcal{E}_{un}(d_j, q_\omega^2)),$$

Then, the issue of system for any user query q is a fuzzy subset of documents characterized by the linguistic membership function $\mathcal{E}_{un}$:

$$\{(d_1, \mathcal{E}_{un}(d_1, q^k)), \ldots, (d_m, \mathcal{E}_{un}(d_m, q^k))\},$$

$$k \in \{1, 2, 3, 4, 5\}.$$

The documents are shown in decreasing order of $\mathcal{E}_{un}$ and arranged in linguistic relevance classes, in such a way that the maximal number of classes is limited by the cardinality of the unbalanced set of labels chosen for represented the linguistic variable *Relavance*.

5 Concluding Remarks

In this contribution we have presented a linguistic IRS using unbalanced linguistic term sets. In such a way, on the one hand, users can use a higher number of discrimination values to assess the importance assigned to the terms of queries, and on the other hand, the system has also a higher number of discrimination values to assess the relevance assigned to the retrieved documents. To do so, we have developed a methodology to manage unbalanced linguistic information based on the linguistic 2-tuple representation model and the linguistic hierarchical contexts. Additionally, this methodology allows us to improve the performance of IRS by increasing the classification levels of the retrieved documents.

References

[1] B. Arfi. Fuzzy decision making in politics: A linguistic fuzzy-set approach (LFSA). *Political Analysis* **13:1** (2005) 23-56.

[2] G. Bordogna and G. Pasi, An fuzzy linguistic approach generalizing Boolean information retrieval: A model and its evaluation. *Journal of the American Society for Information Science and Technology* **44** (1993) 70-82.

[3] G. Bordogna and G. Pasi, Application of the OWA operators to soften information retrieval systems, in: R.R Yager and J. Kacprzyk, Eds., *The Ordered Weighted Averaging Operators: Theory and Applications* (Kluwer Academic Publishers, 1997) 275-294.

[4] G. Bordogna and G. Pasi, An ordinal information retrieval model. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **9** (2001) 63-76.

[5] O. Cordón, F. Herrera, and I. Zwir, Linguistic Modeling by Hierarchical Systems of Linguistic Rules. *IEEE Transactions on Fuzzy Systems* **10:1** (2002) 2-20.

[6] J. Fodor and M. Roubens, Fuzzy Preference Modelling and Multicriteria Decision Support. *Kluwer Academic Publishers* (1994).

[7] F. Herrera and E. Herrera-Viedma, Linguistic decision analisys: steps for solving decision problems under linguistic information. *Fuzzy Sets and Systems* 115 (2000) 67-82.

[8] F. Herrera and E. Herrera-Viedma, Aggregation operators for linguistic

weighted information. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans* **27** (1997) 646-656.

[9] F. Herrera, E. Herrera-Viedma and J.L. Verdegay, Direct approach processes in group decision making using linguistic OWA operators. *Fuzzy Sets and Systems* **79** (1996) 175-190.

[10] F. Herrera and L. Martínez, A 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Transactions on Fuzzy Systems* **8:6** (2000) 746-752.

[11] F. Herrera and L. Martínez, A model based on linguistic 2-tuples for dealing with multigranularity hierarchical linguistic contexts in multiexpert decision-making. *IEEE Transactions on Systems, Man and Cybernetics. Part B: Cybernetics* **31:2** (2001) 227-234.

[12] E. Herrera-Viedma, Modeling the retrieval process for an information retrieval system using an ordinal fuzzy linguistic approach. *Journal of the American Society for Information Science and Technology* **52:6** (2001) 460-475.

[13] E. Herrera-Viedma, An IR model with ordinal linguistic weighted queries based on two weighting elements. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **9** (2001) 77-88.

[14] E. Herrera-Viedma, L. Martínez, F. Mata and F. Chiclana, A consensus support system model for group decision-making problems with multi-granular linguistic preference relations. *IEEE Transaction on Fuzzy Systems* (2005). To appear.

[15] E. Herrera-Viedma and E. Peis, Evaluating the informative quality of documents in SGML-format using fuzzy linguistic techniques based on computing with words. *Information Processing and Management* **39:2** (2003) 195-213.

[16] E. Herrera-Viedma, A.G. López-Herrera, M. Luque and C. Porcel, A fuzzy linguistic IRS model based on a 2-tuple fuzzy linguistic approach. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* (2005). Submited.

[17] E. Herrera-Viedma, A.G. López-Herrera and C. Porcel, Tuning the matching function for a threshold weighting semantics in a linguistic information retrieval system. *International Journal of Intelligent Systems* **20** (2005) 921-937.

[18] E. Herrera-Viedma, G. Pasi and A.G. López-Herrera, Evaluating the information quality of web sites: a qualitative methodology based on fuzzy computing with words. *Journal of the American Society for Information Science and Technology* (2005). To appear.

[19] J. Kacprzyk and S. Zadrozny, Computing with words for text processing: an approach to the text categorization. *Information Sciences* **176:4** (2006) 415-437.

[20] T. Radecki, Fuzzy set theorical approach to document retrieval. *Information Processing & Management* **15** (1979) 247-260.

[21] G. Salton and M.H. McGill, *Introduction to modern information retrieval.* (New York: McGraw-Hill, 1983).

[22] R.R. Yager, A note on weighted queries in information retrieval system. *Journal of the American Society for Information Science and Technology* **38** (1987) 23-24.

[23] R.R. Yager, On ordered weighted averaging operators in multicriteria decision making. *IEEE Transactions on Systems, Man, and Cybernetics* **18** (1988) 183-190.

[24] R.R. Yager, An Approach to Ordinal Decision Making, *International Journal of Approximate Reasoning*, **12** (1995) 237-261.

Analyses of cognitive processes in context
Analyse des processus cognitifs en contexte

Metaknowledge *Métaconnaissances*

volume I

PROCEEDINGS
Eleventh International Conference

ACTES
Onzième conférence internationale

IPMU

Information Processing and Management of Uncertainty in Knowledge-based Systems

Traitement d'information et gestion d'incertitudes dans les systèmes à base de connaissances

Les Cordeliers • Paris • 2006

EDK
EDITIONS MEDICALES ET SCIENTIFIQUES