A Model of an Information Retrieval System with Unbalanced Fuzzy Linguistic Information

Enrique Herrera-Viedma,[1,*] Antonio Gabriel López-Herrera[2,†]
[1]*Department of Computer Science and Artificial Intelligence, University of Granada, 18071, Granada, Spain*
[2]*Department of Computer Science, University of Jaén, Paraje Las Lagunillas, s/n 23071 Jaén, Spain*

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, to improve the system–user interaction, it seems more adequate to express these linguistic weights and degrees by means of unbalanced linguistic scales, that is, linguistic term sets with different discrimination levels on both sides of the middle linguistic term. In this contribution we present an information retrieval system that accepts weighted queries whose weights are expressed using unbalanced linguistic term sets. Then, the system provides the retrieved documents classified in linguistic relevance classes assessed on unbalanced linguistic term sets. To do so, we propose a methodology to manage unbalanced linguistic information and we use the linguistic 2-tuple model as the representation base of the unbalanced linguistic information. Additionally, the linguistic 2-tuple model allows us to increase the number of relevance classes in the output and also to improve the performance of the information retrieval system.

1. INTRODUCTION

Information retrieval (IR) 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[1]:

(1) *a documentary archive*, which stores the documents and the representation of their information contents (index terms)

*Author to whom all correspondence should be addressed: e-mail: viedma@decsai.ugr.es.
†e-mail: aglopez@ujaen.es.

INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS, VOL. 22, 1197–1214 (2007)
 Published online in Wiley InterScience (www.interscience.wiley.com). • DOI 10.1002/int.20244

(2) *a query subsystem*, which allows users to formulate their queries by means of a query language
(3) *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. This is a very complex task because it presents subjectivity and uncertainty. To achieve this, a possible solution consists of 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.[2–9] So different weighted IRSs based on an *ordinal fuzzy linguistic approach*[10–12] were presented in Refs. 13–17. With such a linguistic approach, the weights are assumed to be 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, conversely, the estimated relevance levels of the documents are supplied in a linguistic form (e.g., linguistic terms such as "relevant" or "very relevant" may be used). The problem is that, when using symmetrically and uniformly distributed linguistic term sets, we find the same discrimination levels on both sides of the middle 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 middle label much more than on the left. Similarly, usually users are interested in the relevant documents much more than in the nonrelevant 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 middle linguistic term is assumed. Therefore, in IR, the use of *unbalanced linguistic term sets* (see Figure 1), that is, linguistic term sets with different discrimination levels on both sides of the middle linguistic term, to express weighted queries and the relevance of documents seems more appropriate.

The aim of this contribution is to present a linguistic IRS that manages unbalanced linguistic information to represent the weights of queries and the relevance degrees of retrieved documents. To achieve this, we use hierarchical linguistic contexts based on the linguistic 2-tuple computational model[18,19] and we propose a methodology to manage unbalanced linguistic information. In such a way, we present a linguistic IRS that improves the expressiveness in the system–user interaction process. Furthermore, the use of the 2-tuple model improves the

Figure 1. Example of an unbalanced linguistic term set of seven labels.

performance of the ordinal linguistic IRS because it allows us to represent more classification levels of relevance.

This article is structured as follows. Section 2 introduces the methodology designed to manage unbalanced linguistic information. Section 3 presents the linguistic IRS based on unbalanced linguistic information. In Section 4, an example of the performance of the linguistic IRS proposed is presented. And finally, some concluding remarks are made.

2. A METHODOLOGY TO MANAGE UNBALANCED LINGUISTIC INFORMATION

In this section, we present a methodology to manage unbalanced linguistic information defined using the hierarchical linguistic contexts based on the linguistic 2-tuple computational model.[18,19]

2.1. Linguistic Computational Model Based on 2-tuples

Herrera and Martínez[18] presented a linguistic computational model based on linguistic 2-tuples that carries out processes of computing with words (CW) in a precise way when the linguistic term sets are symmetrically and uniformly distributed. This model is based on the concept of *symbolic translation*. It represents the linguistic information by means of linguistic 2-tuples and defines a set of functions to facilitate computational processes over 2-tuples.

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

$$\Delta : [0, g] \to \mathcal{S} \times [-0.5, 0.5)$$

$$\Delta(\beta) = (s_i, \alpha) \begin{cases} s_i & i = round(\beta) \\ \alpha = \beta - i & \alpha \in [-0.5, 0.5) \end{cases}$$

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

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

Proof. It is trivial; we consider the following function:

$$\Delta^{-1} : \mathcal{S} \times [-0.5, 0.5) \to [0, g]$$

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

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 the value of the symbolic translation: $s_i \in \mathcal{S} \Rightarrow (s_i, 0)$.

The 2-tuples linguistic computational model presents different techniques to manage the linguistic information[18]:

- *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-tuples, with each one representing a counting of information:
  - if $k < l$, then (s_k, α_1) is smaller than (s_l, α_2)
  - if $k = l$, then
    1. if $\alpha_1 = \alpha_2$, then (s_k, α_1), (s_l, α_2) represent the same information
    2. if $\alpha_1 < \alpha_2$, then (s_k, α_1) is smaller than (s_l, α_2)
    3. if $\alpha_1 > \alpha_2$, then (s_k, α_1) is bigger than (s_l, α_2)
- *Negation of 2-tuple* is defined as

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

- *Aggregation of 2-tuples*. Using the functions Δ and Δ^{-1}, any aggregation operator can be easily extended for dealing with linguistic 2-tuples. Some examples are presented in Ref. 18.

2.2. Hierarchical Linguistic Contexts Based on 2-tuples

The hierarchical linguistic contexts were introduced in Refs. 19 and 20 to improve the precision of processes of CW in multigranular linguistic contexts. In this contribution, we use them to manage unbalanced linguistic term sets.

A *linguistic hierarchy* is a set of levels in which each level represents a linguistic term set with different granularity from the remaining levels. Each level is denoted as $l(t, n(t))$, where

(1) t is a number that indicates the level of the hierarchy, and
(2) $n(t)$ is the granularity of the linguistic term set of the level t.

We assume levels containing linguistic terms whose membership functions are triangular and 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, that is, 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, *LH*, as the union of all levels t:

$$LH = \bigcup_t l(t, n(t))$$

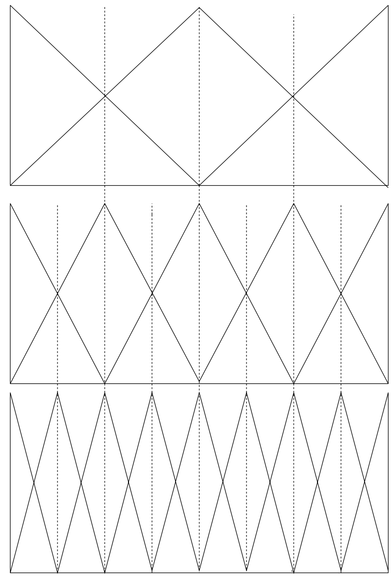

Figure 2. Linguistic hierarchy of three, five, and nine labels.

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

$$\mathcal{S}^{n(t)} = \{s_0^{n(t)}, \ldots, s_{n(t)-1}^{n(t)}\}$$

Generically, we can say that the linguistic term set of level $t+1$ is obtained from its predecessor as

$$l(t, n(t)) \rightarrow l(t+1, 2 \cdot n(t) - 1)$$

A graphical example of a linguistic hierarchy is shown in Figure 2.

In Ref. 19, transformation functions between labels of different levels were developed to make processes of CW without loss of information.

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

$$TF_{t'}^{t} : l(t, n(t)) \rightarrow l(t', n(t'))$$

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

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

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

2.3. A Management Model of Unbalanced Linguistic Information

Here, we propose a model 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 *LH*, carrying out computational operations of unbalanced linguistic information using the 2-tuple computational model.

This management model of unbalanced linguistic information presents two components: a representation model of unbalanced linguistic information and a computational model of unbalanced linguistic information.

2.3.1. A Representation Model of the Unbalanced Linguistic Term Set $\mathcal{S}_{un}$ by Means of a Linguistic Hierarchy LH

To do this, we use different levels of the linguistic hierarchy *LH* to represent both sides of the middle linguistic term. So, the side with more linguistic terms needs a more granular level $l(i, n(i))$ of *LH* and the side with fewer linguistic terms needs a less granular level $l(j, n(j))$ of *LH*, where $i > j$. Concretely, the steps areas follows:

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

Assuming the unbalanced linguistic term set $\mathcal{S}_{un} = \{N = NONE, L = LOW, M = MEDIUM, H = HIGH, QH = QUITE\ HIGH, VH = VERY\ HIGH, T = TOTAL\}$ shown in Figure 1 and the linguistic hierarchy *LH* shown in Figure 2, in Figure 3, an example of linguistic hierarchy is shown in which different levels are used to represent the terms of both sides of the middle term. So, to represent the terms $\{N, L, M\}$, level $l(2,5)$ is used ($t^- = l(2,5)$), and to represent $\{H, QH, VH, T\}$, level $l(3,9)$ is more adequate ($t^+ = l(3,9)$).

2.3.2. An Unbalanced Linguistic Computational Model

To manage unbalanced linguistic information we need a set of computation tools.

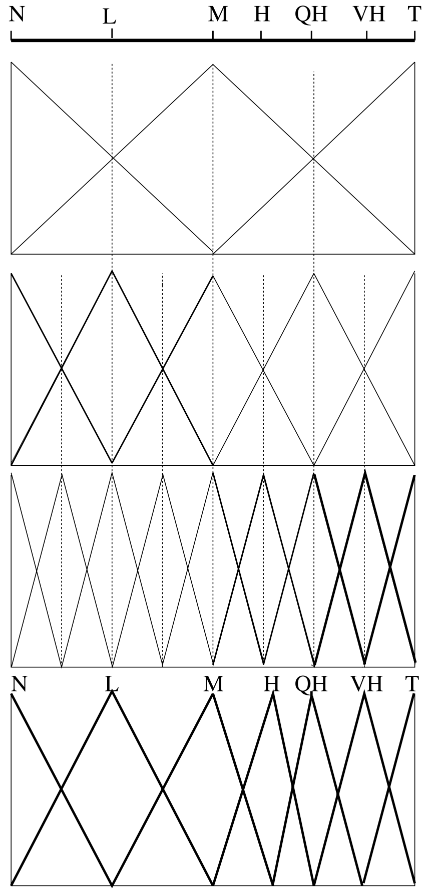

Figure 3. Representation for an unbalanced term set of seven labels.

Previously to carry out any computation task of unbalanced linguistic information, we chose a level $t' \in \{t^-, t^+\}$, such that $n(t') = max\{n(t^-), n(t^+)\}$. Then, we could define three kinds of computation tools.

(1) Comparison of two unbalanced linguistic 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 $TF_{t'}^{t}(s_k^{n(t)}, \alpha_1) = (s_v^{n(t')}, \beta_1)$ and $TF_{t'}^{t}(s_l^{n(t)}, \alpha_2) = (s_w^{n(t')}, \beta_2)$. Then we have
- if $v < w$, then $(s_v^{n(t')}, \beta_1)$ is smaller than $(s_w^{n(t')}, \beta_2)$
- if $v = w$, then
  (a) if $\beta_1 = \beta_2$, then $(s_v^{n(t')}, \beta_1)$, $(s_w^{n(t')}, \beta_2)$ represent the same information
  (b) if $\beta_1 < \beta_2$, then $(s_v^{n(t')}, \beta_1)$ is smaller than $(s_w^{n(t')}, \beta_2)$
  (c) if $\beta_1 > \beta_2$, then $(s_v^{n(t')}, \beta_1)$ is bigger than $(s_w^{n(t')}, \beta_2)$

We should point out that using the comparison of unbalanced 2-tuples we can easily define the comparison operators MAX_{un} and MIN_{un}.

(2) Negation operator of unbalanced linguistic information ($\mathcal{NEG}$). Let $(s_k^{n(t)}, \alpha)$, $t \in \{t^-, t^+\}$ be an unbalanced 2-tuple; then

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

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

(3) 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 a result is obtained, 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 $\mathcal{S}_{un}$.

For example, we can easily define the $LOWA_{un}$ operator, which is an extension of the Linguistic Ordered Weighted Averaging (LOWA) defined in Ref. 11 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*

$$\begin{aligned}\phi_{un}((a_1, \alpha_1), \ldots, (a_m, \alpha_m)) &= W \cdot B^T \\ &= C_{un}^m\{w_w, b_w, k = 1, \ldots, m\} \\ &= w_1 \otimes b_1 \oplus (1 - w_1) \otimes C_{un}^{m-1}\{\beta_h, b_h, h = 2, \ldots, m\}\end{aligned}$$

where $b_i = (a_i, \alpha_i) \in (\mathcal{S} \times [-0.5, 0.5)), W = [w_1, \ldots, w_m]$ is a weighting vector, such that, $w_i \in [0, 1]$ and $\sum_i w_i = 1$, $\beta_h = w_h / \sum_2^m w_k$, $h = 2, \ldots, m$, and B is the associated ordered unbalanced 2-tuple vector. Each element $b_i \in B$ is the ith 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

$$\begin{aligned}C_{un}^2\{w_l, b_l, l = 1, 2\} &= w_1 \otimes b_j \oplus (1 - w_1) \otimes b_i \\ &= TF_t^{t'}(s_k^{n(t')}, \alpha)\end{aligned}$$

where $(s_k^{n(t')}, \alpha) = \Delta(\lambda)$ and $\lambda = \Delta^{-1}(TF_{t'}^{t}(b_i)) + w_1 \cdot (\Delta^{-1}(TF_{t'}^{t}(b_j)) - \Delta^{-1}(TF_{t'}^{t} \times (b_i)))$, $b_j, b_i \in (\mathcal{S} \times [-0.5, 0.5))$, $(b_j \geq b_i)$, $\lambda \in [0, n(t') - 1], t \in \{t^-, t^+\}$.

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[21]:

(1) The 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[22]:

(i) *the classical MIN operator:*

$$LC_1^{\rightarrow}(\omega, a) = MIN_{un}(\omega, a)$$

(ii) *the nilpotent MIN operator*:

$$LC_2^{\rightarrow}(\omega, a) = \begin{cases} MIN_{un}(\omega, a) & \text{if } \omega > \mathcal{NEG}(a) \\ s_0 & \text{otherwise} \end{cases}$$

(iii) *the weakest conjunction*:

$$LC_3^{\rightarrow}(\omega, a) = \begin{cases} MIN_{un}(\omega, a) & \text{if } MAX_{un}(\omega, a) = s_T \\ s_0 & \text{otherwise} \end{cases}$$

where $\omega, a \in \mathcal{S}_{un}$, a is the assessment to aggregate and ω is the associated weight to a.

(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[22]:

(i) *Kleene–Dienes's implication function*:

$$LI_1^{\rightarrow}(\omega, a) = MAX_{un}(\mathcal{NEG}(\omega), a)$$

(ii) *Gödel's implication function*:

$$LI_2^{\rightarrow}(\omega, a) = \begin{cases} s_T & \text{if } \omega \leq a \\ a & \text{otherwise} \end{cases}$$

(iii) *Fodor's implication function*:

$$LI_3^{\rightarrow}(\omega, a) = \begin{cases} s_T & \text{if } \omega \leq a \\ MAX_{un}(\mathcal{NEG}(\omega), a) & \text{otherwise} \end{cases}$$

where $\omega, a \in \mathcal{S}_{un}$, a is the assessment to aggregate and ω is the associated weight to a.

(2) The aggregation of the transformed weighted information by means of an aggregation operator of nonweighted information *f*. As is known, the choice of *h* depends on *f*.

As an *f* operator, we can use the $LOWA_{un}$ with the transformated weighted degrees by *h*.

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* associated with any vector *W*, which allows us to control its aggregation behavior:

$$orness(W) = \left(\frac{1}{m-1}\right) \cdot \left(\sum_{i=1}^{m} (m-i) \cdot w_i\right)$$

Given a weighting vector W, then the closer an OWA operator is to an "or," the closer its orness measure is to 1, whereas the nearer it is to an "and," the closer is the latter measure to 0. We use this good property in our linguistic IRS to evaluate the logical connectives of Boolean queries OR and AND.

3. THE IRS WITH UNBALANCED LINGUISTIC INFORMATION

In this section we present a linguistic IRS that uses an unbalanced linguistic term set $\mathcal{S}$ to express the linguistic assessments in the retrieval process. Particularly, $\mathcal{S}$ presents a higher number of discrimination levels on the right of the middle linguistic term than on the left (e.g., as happens in the example in Figure 1). Then, this IRS accepts linguistically weighted queries and provides linguistic retrieval status values (RSVs) assessed on $\mathcal{S}$ and $\mathcal{S} \times [-0.5, 0.5)$, respectively. The components of this IRS are presented in the following subsections.

3.1. Documentary Archive

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. $\mathcal{F}(d_j, t_i) = 0$ implies that the document d_j is not at all about the concept(s) represented by index term t_i and $\mathcal{F}(d_j, t_i) = 1$ implies that the document d_j is perfectly represented by the concept(s) indicated by t_i. Then each d_j is represented as $R_{d_j} = \Sigma_{i=1}^{l} \mathcal{F}(d_j, t_i)/t_i$.

We assume that the system uses any of the existing weighting methods[1] to compute $\mathcal{F}$.

3.2. The Query Subsystem

The *query subsystem* presents a weighted Boolean query language to express user information needs. In the queries, the terms can be weighted according to two different semantic possibilities, even simultaneously. These semantics are a threshold semantics and a relative importance semantics. As in Ref. 24, we use the linguistic variable *Importance* to express the linguistic weights associated with the query terms. Thus, we consider a set of unbalanced linguistic values $\mathcal{S}_{un}$.

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 that are connected by the logical operators AND ($\wedge$), OR ($\vee$), and NOT ($\neg$).

Therefore, a query Q is any legitimate Boolean expression whose atomic components (atoms) are 3-tuples $\langle t_i, c_i^1, c_i^2 \rangle$ belonging to the set $\mathcal{T} \times \mathcal{S}_{un}^2$; $t_i \in \mathcal{T}$

and c_i^1 and c_i^2 are linguistic values of the linguistic variable *Importance* modeling 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 Q of the legitimate queries is defined by the following syntactic rules:

(1) $\forall q = \langle t_i, c_i^1, c_i^2 \rangle \in \mathcal{T} \times \mathcal{S}_{un}^2 \rightarrow q \in Q$
(2) $\forall q, p \in Q \rightarrow q \wedge p \in Q$
(3) $\forall q, p \in Q \rightarrow q \vee p \in Q$
(4) $\forall q \in Q \rightarrow \neg q \in Q$
(5) All legitimate queries $p \in Q$ are only those obtained by applying rules 1–4, inclusive.

3.3. 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 one weighted term is evaluated by means of a constructive bottom-up process that 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. According to the threshold semantics, 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. To model the interpretation of the threshold semantics, we use the matching function described in Ref. 25 but, defined in a 2-tuple linguistic context, it is called g_{un}, and defined as $g_{un} : \mathcal{D} \times \mathcal{T} \times \mathcal{S}_{un} \rightarrow \mathcal{S}_{un} \times [-0.5, 0.5)$. Then, given an atom $\langle t_i, c_i^1, c_i^2 \rangle$, $t_i \in \mathcal{T}$, 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:

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

where $(s_a, \alpha_a) = \Delta((n(t) - 1) \cdot \mathcal{F}(d_j, t_i))$, $\Delta : [0, n(t) - 1] \rightarrow \mathcal{S}_{un} \times [-0.5, 0.5)$, with $t = t^-$ if $\mathcal{F}(d_j, t_i) \leq 0.5$ and $t = t^+$ if $\mathcal{F}(d_j, t_i) > 0.5$, t^- and t^+ being the levels of *LH*.

3. *Evaluation of subexpressions and modeling 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 that 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 $RSV_j^{i,1} \in (\mathcal{S}_{un} \times [-0.5, 0.5))$ with respect to each atom $\langle t_i, c_i^1, c_i^2 \rangle$ 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 that should guarantee that the more important the query terms, the more important they are in the determination of the RSVs.

Yager[26] 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 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 very restrictive and inclusive behaviors, respectively. This fact means that the retrieval process may be deceptive because, on the one hand, the linguistic MIN t-norm may cause the rejection of useful documents due to 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 due to 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 $LC_1^{\rightarrow}$ and $LI_1^{\rightarrow}$, 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 their 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 $orness(W) \leq 0.5$ to model the AND connectives and an unbalanced operator ϕ_{un}^2 with $orness(W) > 0.5$ 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 $h = \phi_{un}^1$ and $f = MAX_{un}(\mathcal{NEG}(weight, 0)$, unbalanced value) and, if it is disjunctive, then we use $h = \phi_{un}^2$; then $f = MIN_{un}((weight, 0)$, unbalanced value).

Briefly, given a document d_j, we evaluate its relevance with respect to a subexpression q_v, called $RSV_j^v \in (\mathcal{S} \times [-0.5, 0.5))$ as follows:

(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, $RSV_j \in (\mathcal{S}_{un} \times [-0.5, 0.5))$ is given as follows:

(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.

Remark 2: On the NOT Operator. We should note that, if a query is in CNF or DNF, we have to define the negation operator only at the level of single atoms. This simplifies the definition of the NOT operator. As was done in Ref. 15, the evaluation of document d_j for a negated weighted atom $\langle \neg t_i, c_i^1, c_i^2 \rangle$ is obtained from the negation of the index term weight $\mathcal{F}(t_i, d_j)$. This means to calculate the threshold matching function g_{un} from the linguistic unbalanced value $\mathcal{NEG}(\Delta((n(t)-1) \cdot \mathcal{F}(d_j, t_i)))$, with $t = t^-$ if $\mathcal{F}(d_j, t_i) \leq 0.5$ and $t = t^+$ if $\mathcal{F}(d_j, t_i) > 0.5$.

Briefly, this evaluation subsystem can be synthesized by means of a general linguistic evaluation function $E_{un}: \mathcal{D} \times Q \to \mathcal{S}_{un} \times [-0.5, 0.5)$, which evaluates the different kinds of preprocessed queries, $\{q = \langle t_i, c_i^1, c_i^2 \rangle, q \wedge p, q \vee p\}$ according to the following five rules:

(1) *Atoms*: $E_{un}(d_j, q^1) = g_{un}(d_j, t_i, c_i^1)$, such that, $q^1 = \langle t_i, c_i^1, c_i^2 \rangle$.
(2) *Conjunctive subexpressions*:

$$E_{un}(d_j, q^2) = \phi_{un}^1(MAX_{un}(\mathcal{NEG}(c_1^2,0), E_{un}(d_j, q_1^1)), \ldots, MAX_{un}(\mathcal{NEG}(c_\eta^2,0), E_{un}(d_j, q_\eta^1)))$$

where η is the number of atoms of q^2.

(3) *Disjunctive subexpressions*:

$$E_{un}(d_j, q^3) = \phi_{un}^2(MIN_{un}((c_1^2,0), E_{un}(d_j, q_1^1)), \ldots, MIN_{un}((c_\eta^2,0), E_{un}(d_j, q_\eta^1)))$$

where η is the number of atoms of q^3.

(4) *Query in CNF*:

$$E_{un}(d_j, q^4) = \phi^1_{un}(E_{un}(d_j, q^3_1), \ldots, E_{un}(d_j, q^3_\omega))$$

where ω is the number of disjunctive subexpressions.

(5) *Query in DNF*:

$$E_{un}(d_j, q^5) = \phi^1_{un}(E_{un}(d_j, q^2_1), \ldots, E_{un}(d_j, q^2_\omega))$$

where ω is the number of conjunctive subexpressions.

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

$$\{(d_1, E_{un}(d_1, q^k)), \ldots, (d_m, E_{un}(d_m, q^k))\}, k \in \{1, 2, 3, 4, 5\}.$$

The documents are shown in decreasing order of 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 to represent the linguistic variable *Relavance*.

4. OPERATION OF THE IRS PROPOSED USING UNBALANCED LINGUISTIC INFORMATION

In this section, we present an example of the performance of the proposed IRS model using unbalanced linguistic information.

Let us suppose a small database containing a set of seven documents $\mathcal{D} = \{d_1, \ldots, d_7\}$, represented by means of a set of 10 index terms $\mathcal{T} = \{t_1, \ldots, t_{10}\}$. Documents are indexed by means of a numeric indexing function $\mathcal{F}$, which represents them as follows:

$$d_1 = 0.7/t_5 + 0.4/t_6 + 1/t_7$$

$$d_2 = 1/t_4 + 0.6/t_5 + 0.8/t_6 + 0.9/t_7$$

$$d_3 = 0.5/t_2 + 1/t_3 + 0.8/t_4$$

$$d_4 = 0.9/t_4 + 0.5/t_6 + 1/t_7$$

$$d_5 = 0.7/t_3 + 1/t_4 + 0.4/t_5 + 0.8/t_9 + 0.6/t_{10}$$

$$d_6 = 0.8/t_5 + 0.99/t_6 + 0.8/t_7$$

$$d_7 = 0.8/t_5 + 0.02/t_6 + 0.8/t_7 + 0.9/t_8$$

We use the set of the seven unbalanced labels given in the example of Figure 1 and its linguistic hierarchy shown in Figure 3, which has two levels: $LH = l(1,5) \bigcup l(2,9)$, where

- $l(1,5) = \{N, L, M, H, T\}$ and
- $l(2,9) = \{N, VL, QL, L, M, H, QH, VH, T\}$.

So, we represent these documents using a 2-tuple linguistic representation applying the function Δ over index term weights $\mathcal{F}(d_j, t_i)$ and a transformation $TF_t^{t'}$, $t \in \{t^-, t^+\}$, where $t^- = l(1,5)$ and $t^+ = l(2,9)$:

$$d_1 = (QH, -0.4)/t_5 + (M, -0.4)/t_6 + (T, 0.0)/t_7$$

$$d_2 = (T, 0.0)/t_4 + (H, -0.2)/t_5 + (QH, 0.4)/t_6 + (VH, 0.2)/t_7$$

$$d_3 = (M, 0.0)/t_2 + (T, 0.0)/t_3 + (QH, 0.4)/t_4$$

$$d_4 = (VH, 0.2)/t_4 + (M, 0.0)/t_6 + (T, 0.0)/t_7$$

$$d_5 = (QH, -0.4)/t_3 + (T, 0.0)/t_4 + (M, -0.4)/t_5 + (QH, 0.4)/t_9 + (H, -0.2)/t_{10}$$

$$d_6 = (QH, 0.4)/t_5 + (T, -0.08)/t_6 + (QH, 0.4)/t_7$$

$$d_7 = (QH, 0.4)/t_5 + (N, 0.08)/t_6 + (QH, 0.4)/t_7 + (VH, 0.2)/t_8$$

Then, we consider that a user formulates the following query: $q = ((t_5, QH, VH) \wedge (t_6, L, L)) \vee (t_7, H, L)$. Then, its evaluation is as follows:

1. *Preprocessing of the query:* The query q is in DNF, but it presents one subexpression with only one atom. Therefore, q must be preprocessed and transformed into a normal form with every one of its subexpressions with a minimum number of two atoms. Then, q is transformed into the following equivalent query: $q' = ((t_5, QH, VH) \vee (t_7, H, L)) \wedge ((t_6, L, L) \vee (t_7, H, L))$, which is expressed in CNF.

2. *Evaluation of the atoms with respect to the threshold semantics:* After the query q is transformed into normal form, we evaluate atoms according to the threshold semantics by means of g_{un} and we obtain the following:

- For t_5:

$$\{RSV_6^{5,1} = (QH, 0.4), RSV_7^{5,1} = (QH, 0.4)\}$$

- For t_6:

$$\{RSV_1^{6,1} = (M, -0.4), RSV_2^{6,1} = (QH, 0.4), RSV_4^{6,1} = (M, 0.0), RSV_6^{6,1} = (T, -0.08)\}$$

- For t_7:

$$\{RSV_1^{7,1} = (T, 0.0), RSV_2^{7,1} = (VH, 0.2), RSV_4^{7,1} = (T, 0.0), RSV_6^{7,1} = (QH, 0.4),$$
$$RSV_7^{7,1} = (QH, 0.4)\}$$

where, for example, the $RSV_2^{7,1}$ is calculated as follows:

$$RSV_2^{7,1} = g_{un}(d_2, t_7, H) = (VH, 0.2)$$

such that $(VH, 0.2) \geq (H, 0.0)$ and $(VH, 0.2) = \Delta(8 \cdot 0.9)$, with $8 = n(t^+) - 1$ and $0.9 = \mathcal{F}(d_2, t_7)$.

3. *Evaluation of subexpressions and modeling the relative importance semantics:* The query q' has two subexpressions, and both present two atoms, $q_1' = (t_5, QH, VH) \vee (t_7, H, L)$ and $q_2' = (t_6, L, L) \vee (t_7, H, L)$. Each subexpression is in disjunctive form, and, thus, we must use an unbalanced $LOWA_{un}$ operator ϕ_{un}^2 with orness measure $orness(W) > 0.5$ (e.g., with $(W = [0.8, 0.2])$) together with the transformation function $MIN_{un}(weight,$ unbalanced value) to evaluate them. Then, the results of evaluation applying the relative importance semantics are

- For q_1':

$$\{RSV_1^1 = (L, -0.2), RSV_2^1 = (L, -0.2), RSV_4^1 = (L, -0.2), RSV_6^1 = (QH, -0.48), RSV_7^1 = (QH, -0.48)\}$$

- For q_2':

$$\{RSV_1^2 = (L, 0.0), RSV_2^2 = (L, 0.0), RSV_4^2 = (L, 0.0), RSV_6^2 = (L, 0.0), RSV_7^2 = (L, -0.2)\}$$

where, for example, the RSV_6^1 is calculated as follows:

$$\begin{aligned}
RSV_6^1 &= \phi_{un}^2(MIN_{un}((c_5^2, 0), RSV_6^{5,1}), MIN_{un}((c_7^2, 0), RSV_6^{7,1}))\\
&= \phi_{un}^2(MIN_{un}((VH, 0), (QH, 0.4)), MIN_{un}(L, 0), (QH, 0.4))\\
&= \phi_{un}^2((QH, 0.4), (L, 0.0))\\
&= \Delta^{-1}(TF_{t'}^{t^+}(QH, 0.4)) \cdot 0.8 + \Delta^{-1}(TF_{t'}^{t^-}(L, 0.0)) \cdot 0.2\\
&= \Delta^{-1}(QH, 0.4) \cdot 0.8 + \Delta^{-1}(QL, 0.0) \cdot 0.2 = \Delta(5.52)\\
&= (QH, -0.48) \Rightarrow TF_{t^+}^{t'}(QH, -0.48) = (QH, -0.48)
\end{aligned}$$

such that $t^+ = t'$.

4. *Evaluation of the whole query:* We obtain the evaluation of the whole query using an unbalanced $LOWA_{un}$ operator ϕ_{un}^1 with $orness(W) < 0.5$ (e.g., with $(W = [0.2, 0.8])$):

$$\{RSV_6 = (L, 0.352), RSV_7 = (L, 0.192), RSV_1 = (L, -0.16), RSV_2 = (L, -0.16), RSV_4 = (L, -0.16)\}$$

The best retrieved document is d_6, which is calculated as

$$
\begin{aligned}
RSV_6 &= \phi_{un}^2(RSV_6^1, RSV_6^2) \\
&= \Delta^{-1}(TF_{t'}^{t^+}(QH, -0.48))\cdot 0.2 + \Delta^{-1}(TF_{t'}^{t^-}(L, 0.0))\cdot 0.8 \\
&= \Delta^{-1}(QH, -0.48)\cdot 0.2 + \Delta^{-1}(QL, 0.0)\cdot 0.8 = \Delta(2.704) \\
&= (L, -0.296) \Rightarrow TF_{t^-}^{t'}(L, -0.296) = (L, 0.352)
\end{aligned}
$$

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 achieve this, 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 the IRS by increasing the classification levels of the retrieved documents.

References

1. Salton G, McGill MH. Introduction to modern information retrieval. New York: McGraw-Hill; 1983.
2. Herrera-Viedma E, Martínez L, Mata F, Chiclana F. A consensus support system model for group decision-making problems with multi-granular linguistic preference relations. IEEE Trans Fuzzy Syst 2005;13:644–658.
3. Herrera-Viedma E, Peis E. Evaluating the informative quality of documents in SGML-format using fuzzy linguistic techniques based on computing with words. Inform Process Manag 2003;39:195–213.
4. Herrera-Viedma E, López-Herrera AG, Porcel C. Tuning the matching function for a threshold weighting semantics in a linguistic information retrieval system. Int J Intell Syst 2005;20:921–937.
5. Herrera-Viedma E, Pasi G, López-Herrera AG, Porcel C. Evaluating the information quality of web sites: A methodology based on fuzzy computing with words. J Am Soc Inform Sci Technol 2006;57:538–549.
6. Herrera-Viedma E, López-Herrera AG, Luque M, Porcel C. A fuzzy linguistic IRS model based on a 2-tuple fuzzy linguistic approach. Int J Uncertainty, Fuzziness Knowl Based Syst 2007;15:225–250.
7. Mauris G, Galichet S, Dayre E. Predicting the gear ratio changing time by a fuzzy linguistic approach—An application to truck driving. In: Proc 19th IEEE Instrumentation and Measurement Technology Conference 2 (IMTC/2002); 2002. pp 1609–1612.
8. Valet L, Mauris G, Bolon P, Keskes N. A fuzzy linguistic-based software tool for seismic image interpretation. IEEE Trans Instrum Meas 2003;52:675–680.
9. Xu Z. An approach based on the uncertain LOWG and induced uncertain LOWG operators to group decision making with uncertain multiplicative linguistic preference relations. Decis Support Syst 2005;41:488–499.

10. Herrera F, Herrera-Viedma E. Linguistic decision analysis: Steps for solving decision problems under linguistic information. Fuzzy Set Syst 2000;115:67–82.
11. Herrera F, Herrera-Viedma E, Verdegay JL. Direct approach processes in group decision making using linguistic OWA operators. Fuzzy Set Syst 1996;79:175–190.
12. Herrera F, Herrera-Viedma E, Verdegay JL. A linguistic decision process in group decision making. Group Decis Negot 1996;5:165–176.
13. Bordogna G, Pasi G. Application of the OWA operators to soften information retrieval systems. In: Yager RR, Kacprzyk J, editors. The ordered weighted averaging operators: Theory and applications. Dordrecht: Kluwer Academic Publishers; 1997. pp 275–294.
14. Bordogna G, Pasi G. An ordinal information retrieval model. Int J Uncertainty Fuzziness Knowl Based Syst 2001;9:63–76.
15. Herrera-Viedma E. Modeling the retrieval process for an information retrieval system using an ordinal fuzzy linguistic approach. J Am Soc Inform Sci Technol 2001;52:460–475.
16. Herrera-Viedma E. An IR model with ordinal linguistic weighted queries based on two weighting elements. Int J Uncertainty Fuzziness Knowl Based Syst 2001;9:77–88.
17. Herrera-Viedma E, Cordón O, Luque M, López-Herrera AG, Muñoz AM. A model of fuzzy linguistic IRS based on multi-granular linguistic information. Int J Approx Reason 2003; 34:221–239.
18. Herrera F, Martínez L. A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Trans Fuzzy Syst 2000;8:746–752.
19. Herrera F, Martínez L. A model based on linguistic 2-tuples for dealing with multigranularity hierarchical linguistic contexts in multiexpert decision-making. IEEE Trans Syst, Man Cybern B Cybern 2001;31:227–234.
20. Cordón O, Herrera F, Zwir I. Linguistic modeling by hierarchical systems of linguistic rules. IEEE Trans Fuzzy Syst 2002;10:2–20.
21. Herrera F, Herrera-Viedma E. Aggregation operators for linguistic weighted information. IEEE Trans Syst Man Cybern A: Syst Hum 1997;27:646–656.
22. Fodor J, Roubens M. Theory and decision library series D: Fuzzy preference modeling and multicriteria decision support. London: Kluwer; 2002.
23. Yager RR. On ordered weighted averaging operators in multicriteria decision making. IEEE Trans Syst Man Cybern 1988;18:183–190.
24. Bordogna G, Pasi G. An fuzzy linguistic approach generalizing Boolean information retrieval: A model and its evaluation. J Am Soc Inform Sci Technol 1993;44:70–82.
25. Radecki T. Fuzzy set theorical approach to document retrieval. Inform Process Manag 1979;15:247–260.
26. Yager RR. A note on weighted queries in information retrieval system. J Am Soc Inform Sci Technol 1987;38:23–24.