A New Model of Linguistic Weighted Information Retrieval System[1]

E. Herrera-Viedma
Dept. of Computer Science and Artificial Intelligence, Library Science Studies School, University of Granada, 18017 – Granada
viedma@decsai.ugr.es

A. G. López-Herrera
Dept. of Computer Science and Artificial Intelligence, Library Science Studies School, University of Granada, 18017 – Granada
agabriel@ugr.es

C. Porcel
Dept. of Computer Science and Artificial Intelligence, Library Science Studies School, University of Granada, 18017 – Granada
cporcel@invest.ugr.es

Abstract

In this paper a new modelling for a weighted Information Retrieval System (IRS) in a linguistic context is proposed. This linguistic IRS (LIRS) achieves more precise and consistent relevance degrees that early weighted IRSs proposed [8, 9]. To do this, a new redefinition of matching function defined in [11] is used.

Keywords: Fuzzy Information Retrieval, Linguistic Modelling, Weighted Queries.

1 Introduction

The main activity of an Information Retrieval System (IRS) is the gathering of pertinent archived documents that better satisfy the user queries. IRSs present three components to carry out this activity [8, 9]:

i) a *database*: to store the documents (***D***) and the index terms (***T***),
ii) a *query subsystem*: to formulate the user queries,
iii) an *evaluation subsystem:* to obtain the Retrieval Status Values (RSVs) for each document.

The query subsystem supports the user-IRS interaction, and therefore, it should be able to deal with the imprecision and vagueness typical of human communication. This aspect may be modelled by means of the introduction of weights in the query language. Many authors have proposed weighted IRS models using Fuzzy Set Theory [1, 2, 5, 6], in which they assume numeric weights. On the other hand, some fuzzy linguistic IRS models [3, 4, 8, 9, 10, 15] have been proposed using a *fuzzy linguistic approach* [18] to model the query weights and document scores. A useful fuzzy linguistic approach which allows us to reduce the complexity of the design for the IRSs [8, 9] is called the *ordinal fuzzy linguistic approach* [13, 14].

In any weighted IRS we have to establish the semantics associated with the query. There are four semantic possibilities [2, 8, 15]:

i) weights as a measure of the importance of a specific element in representing the query,
ii) as a threshold to aid in matching a specific document to the query,
iii) as a description of an ideal or perfect document, and
iv) as a limit on the amount of documents to be retrieved for a specific element.

In this contribution we present a new modelling of a linguistic IRS. It softens the behaviour of that defined in [8] and allows achieving more precise RSVs. To do that, we use the 2-tuple symmetrical matching function defined in [11]. Then with the 2-tuple fuzzy linguistic representation model [12] we can improve the precision in the representation of linguistic information and with the 2-tuple computational model we can avoid the loss of information in the computation of the linguistic RSVs.

So, the paper is structured as follows. Section 2 presents the 2-tuple fuzzy linguistic approach. Section 3 provides overview of the 2-tuple linguistic

[1] This work has been supported by the Research Project TIC2003-07977.

symmetrical matching function. Section 4 presents the new LIRS proposed and accomplishes a study of its performance. And finally, in Section 5, some concluding remarks are pointed out.

2 A 2-tuple fuzzy linguistic approach

The *ordinal fuzzy linguistic approach* is an approximate technique appropriate to deal with qualitative aspects of problems. An ordinal fuzzy linguistic approach is defined by considering a finite and totally ordered label set $S = \{s_0, ..., s_T\}$, $T+1$ is the cardinality of S in the usual sense, and with odd cardinality (7 or 9 labels) (**Example 1:** $S= \{s_0 =$*Null (N),* $s_1=$*Extremely_Low (EL),* $s_2=$*Very_Low (VL),* $s_3=$*Low (L),* $s_4=$*Medium (M),* $s_5=$*High (H),* $s_6=$*Very_High (VH),* $s_7=$*Extremely_High (EH),* $s_8=$*Total (TO)*$\}$). The mid term representing an assessment of "approximately 0.5" and the rest of the terms being placed symmetrically around it. The semantics of the linguistic terms set is established from the ordered structure of the terms set by considering that each linguistic term for the pair (s_i, s_{T-i}) is equally informative. For each label s_i is given a fuzzy number defined on the [0, 1] interval, which is described by a membership function. The computational model to combine ordinal linguistic information is based on the symbolic approach. It presents the following limitation [12]. Let S be a linguistic term set, if a symbolic method aggregating linguistic information obtains a value $\beta \in [0, T]$, and $\beta \notin \{0, \ldots, T\}$ then an approximation function (app(.)) is used to express the index of the result in S [12]. For example, in the LOWA, app(.) is the simple function *round* [14].

Definition 1. [12] Let $\beta \in [0, T]$ be the result of an aggregation of the indexes of a set of labels assessed in a linguistic term set S, i.e., the result of a symbolic aggregation operation. Let $i = round(\beta)$ and $\alpha_i = \beta - i$ be two values, such that, $i \in \{0, 1, .., T\}$ and $\alpha_i \in [-.5, .5)$ then α_i is called a *Symbolic Translation*.

From this concept, F. Herrera and L. Martínez developed a linguistic representation model which represents the linguistic information by means of 2-tuples (s_i, α_i), $s_i \in S$ and $\alpha_i \in [-.5, .5)$ [12]; where s_i represents the linguistic label of the information, and α_i is a numerical value expressing the value of the translation from the original result β to the closest index label i in S.

This model defines a set of transformation functions between numeric values and linguistic 2-tuples.

Definition 2. [12] Let S be a linguistic term set and $\beta \in [0, T]$, then the 2-tuple that expresses the equivalent information to β is obtained with the following function: $\Delta:[0, T] \rightarrow Sx[-.5, .5)$; $\Delta(\beta) = (s_i, \alpha_i)$, *with* $i=round(\beta)$ *and* $\alpha_i=\beta-i$ $(\alpha_i \in [-.5,.5))$, where s_i has the closest index label to "β" and "α_i" is the value of the symbolic translation.

Proposition 1. [12] Let (s_i, α_i), $s_i \in S$ be a linguistic 2-tuple. There is always a Δ^{-1} function, such that, from a 2-tuple it returns its equivalent numerical value $\beta \in [0, T] \subset \Re$.

Remark 1: [12] From Definition 2 and Proposition 1, it is obvious that the conversion of a linguistic term into a linguistic 2-tuple consists of adding a value 0 as symbolic translation: $s_i \in S \rightarrow (s_i, 0)$.

The 2-tuple linguistic computational model operates with the 2-tuples without loss of information and is based on the following operations [12]:

1 *Negation operator of a 2-tuple*: $Neg(s_i, \alpha_i) = \Delta(T - \Delta^{-1}(s_i, \alpha_i))$.
2 *Comparison of 2-tuples:* The comparison of linguistic information represented by 2-tuples is carried out according to an ordinary lexicographic order.
3 *Aggregation of 2-tuples:* Using the functions Δ and Δ^{-1} any numerical aggregation operator can be easily extended for dealing with linguistic 2-tuples.

Definition 3. [17] Let $A = \{a_1, ..., a_m\}$, $a_k \in [0,1]$ be a set of assessments to aggregated, then the OWA operator, ϕ, is defined as $\phi(a_1, ..., a_m) = W \cdot B^T$, where $W = [w_1, ..., w_m]$, is a weighting vector, such that $w_i \in [0,1]$ and $\Sigma_i w_i = 1$, and $B = \{b_1, ..., b_m\}$ is a vector associated to A, such that, $B = \sigma(A) = \{a_{\sigma(1)}, ..., a_{\sigma(m)}\}$, with σ being a permutation over the set of assessments A, such that $a_{\sigma(j)} \leq a_{\sigma(i)}$ $\forall i \leq j$.

A 2-tuple linguistic extended definition of ϕ would be as follows:

Definition 4. Let $A = \{(a_1, \alpha_1), ..., (a_m, \alpha_m)\}$ be a set of assessments in the linguistic 2-tuple domain, then

the 2-tuple linguistic OWA operator, ϕ_{2t} is defined as $\phi_{2t}((a_1,\alpha_1),...,(a_m,\alpha_m)) = \Delta(W \cdot B^T)$, $B = \sigma(A) = \left\{(\Delta^{-1}(a_1,\alpha_1))_{\sigma(1)},...,(\Delta^{-1}(a_m,\alpha_m))_{\sigma(m)}\right\}$.

3 2-tuple linguistic symmetrical matching function

In [11] we proposed a new matching function to model the symmetrical threshold semantics. This matching function has the following characteristics:

- it is based on the symmetrical matching function proposed in [8], and therefore, it has a symmetric behaviour in both sides of the mid threshold value because it is defined to distinguish two situations in the threshold interpretation: i) when the threshold value (s_b) is on the left of the mid term and ii) when it is on the right. It assumes that a user may use presence weights or absence weights in the formulation of weighted queries. Then, it is symmetrical with respect to the mid threshold value,
- it uses the 2-tuple linguistic representation model and the 2-tuple linguistic computational model for operating with the 2-tuples without loss of information,
- and it softens the behaviour, improves the performance and achieves more consistent and precise relevance degrees.

This new 2-tuple symmetrical matching function is called g_{2t}, and it is defined as: g_{2t}: $\boldsymbol{D}\times\boldsymbol{T}\times(S\times[-.5, .5)) \rightarrow S\times[-.5, .5))$, with

$$g_{2t}(d_j,t_i,(s_b,0)) = \begin{cases} \Delta(\beta_2 + \frac{T}{2}) & if\ (s_a,\alpha_a) \geq (s_b,0) \wedge (s_b,0) \geq (s_{T/2},0) \\ \Delta(\beta_1) & if\ (s_a,\alpha_a) < (s_b,0) \wedge (s_b,0) \geq (s_{T/2},0) \\ \Delta(\beta_2^* + \frac{T}{2}) & if\ (s_a,\alpha_a) \leq (s_b,0) \wedge (s_b,0) < (s_{T/2},0) \\ \Delta(\beta_1^*) & if\ (s_a,\alpha_a) > (s_b,0) \wedge (s_b,0) < (s_{T/2},0) \end{cases} \quad \textbf{(1)}$$

where $\beta_2 = \frac{T\cdot(a_2-u)}{2\cdot(T-u)}$, $\beta_1 = \frac{a_1\cdot T}{2\cdot u}$, $\beta_2^* = \frac{T\cdot(u-a_1)}{2\cdot u}$, $\beta_1^* = \frac{T\cdot(T-a_2)}{2\cdot(T-u)}$, $u = \Delta^{-1}(s_b,0)$, $s_b = a_1 = T\cdot F(d_k, t_i)$, $a_2 = T\cdot F(d_j,t_i)$ and $(s_b,0)$ is the threshold value in 2-tuple form.

4 A new 2-tuple linguistic information retrieval system

The linguistic weighted IRS that we define in this paper presents the following elements to carry out its activity:

1. *Database:* we assume a database of a traditional fuzzy IRS as in [6, 16]. The database stores the finite set of documents $\boldsymbol{D} = \{d_1,...,d_m\}$ represented by a finite set of index terms $\boldsymbol{T} = \{t_1,...,t_l\}$, which describe the subject content of the documents. The representation of a document is a fuzzy set of terms characterized by a numeric indexing function $\boldsymbol{F}: \boldsymbol{D}\times\boldsymbol{T} \rightarrow [0, 1]$, which is called index term weight and it represents the degree of significance of t_i in d_j.

2. *Query subsystem:* we use a query subsystem with a fuzzy linguistic weighted Boolean query language to express user information needs. With this language each query is expressed as a combination of the weighted index terms that are connected by logical operators AND ($\wedge$), OR ($\vee$), and NOT($\neg$). The weights are ordinal linguistic values taken from a label set $\boldsymbol{S}$, and they are associated with a symmetrical threshold semantics [8, 9]. As in [3], our atomic components are pairs, $\langle t_i, c_i\rangle$, where t_i is an index term, but defining the linguistic variable *Importance* with the ordinal linguistic approach and associating c_i with a symmetrical threshold semantics. Accordingly, the set $\boldsymbol{Q}$ of the legitimate queries is defined by the following syntactic rules:

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

3. *Evaluation subsystem:* The evaluation subsystem for weighted Boolean queries acts by means of a constructive bottom-up process based on the *criterion of separability* [7]. The RSVs of the documents are 2-tuple linguistic values whose linguistic components are taken from the linguistic variable *Importance* but representing the concept of *relevance*. Therefore, the set of linguistic terms $\boldsymbol{S}$ is also assumed to represent the relevance values. The evaluation subsystem acts in two steps:
Firstly, the documents are evaluated according to their relevance only to atoms of the query. In this step, the symmetrical threshold semantics is applied

in the evaluation of atoms by means of the new 2-tuple symmetrical matching function g_{2t}.

We should point out that whereas the traditional threshold matching function are always non-decreasing [15], g is non-decreasing on the right of the mid term and decreasing on the left of the mid term in order to be consistent with the meaning of the symmetrical threshold semantics.
Secondly, the documents are evaluated according to their relevance to Boolean combinations of atomic components, and so on, working in a bottom-up fashion until the whole query is processed. In this step, the logical connectives AND and OR are modelled by means of LOWA operators with $orness(W)<0.5$ and $orness(W) \geq 0.5$ respectively, being $orness(W)$ a orness measure introduced by Yager in [17] to classify the aggregation of the OWA operators: $orness(W)=(1/m-1)(\sum_{i=1}^{m} (m-i)\, w_i)$.

Remark 2: We should point out that if we have a negated query, or a negated subexpression, or a negated atom, their evaluation is obtained from the negation of the relevance results computed for the query, or the subexpression, or atom in a no-negated situation.

4.1 An example of operation

In this section, we present an example of performance of the linguistic weighted IRS defined in Section 3.

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

$d_1 = 0.7/t_5 + 0.4/t_6 + 1/t_7$
$d_2 = 1/t_4 + 0.6/t_5 + 0.8/t_6 + 0.9/t_7$
$d_3 = 0.5/t_2 + 1/t_3 + 0.8/t_4$
$d_4 = 0.9/t_4 + 0.5/t_6 + 1/t_7$
$d_5 = 0.7/t_3 + 1/t_4 + 0.4/t_5 + 0.8/t_9 + 0.6/t_{10}$
$d_6 = 0.8/t_5 + 0.99/t_6 + 0.8/t_7$
$d_7 = 0.8/t_5 + 0.02/t_6 + 0.8/t_7 + 0.9/t_8$

Using the set of the nine labels given in Example 1 to provide the linguistic weighted queries, consider that a user formulates the following query:

$$q = ((t_5,VH) \vee (t_7,H)) \wedge ((t_6,L) \vee (t_7,H)).$$

Then, the evaluation process of this query is developed in the following steps:

Evaluation of the atoms with respect to the Symmetrical threshold semantics.

In this step, firstly, we obtain the documents represented in a 2-tuple linguistic form applying the function Δ over index term weights $F(d_j, t_i)$:

$d_1 = (VH,-.4)/t_5 + (L,.2)/t_6 + (TO,0)/t_7$
$d_2 = (TO,0)/t_4 + (H,-.2)/t_5 + (VH,.4)/t_6 + (EH,.2)/t_7$
$d_3 = (M,0)/t_2 + (TO,0)/t_3 + (VH,.4)/t_4$
$d_4 = (EH,.2)/t_4 + (M,0)/t_6 + (TO,0)/t_7$
$d_5 = (VH,-.4)/t_3 + (TO,0)/t_4 + (L,.2)/t_5 + (VH,.4)/t_9 + (H,-.2)/t_{10}$
$d_6 = (VH,.4)/t_5 + (TO,-.08)/t_6 + (VH,.4)/t_7$
$d_7 = (VH,.4)/t_5 + (N,.16)/t_6 + (VH,.4)/t_7 + (EH,.2)/t_8.$

Then, we evaluate atoms according to the symmetrical threshold semantics by means of g_{2t}:

- (t_5,VH) :
$$\{RSV_1^5 = (M,-27),\ RSV_2^5 = (L,.2),$$
$$RSV_5^5 = (VL,.13),\ RSV_6^5 = (H,-.2),$$
$$RSV_7^5 = (H,-.2)\}$$
- (t_6,L):
$$\{RSV_1^6 = (M,-.16),\ RSV_2^6 = (EL,.28),$$
$$RSV_4^6 = (L,.2),\ RSV_6^6 = (N,.06),$$
$$RSV_7^6 = (TO,-.16)\}$$
- (t_7,H):
$$\{RSV_1^7 = (TO,0),\ RSV_2^7 = (EH,-.07),$$
$$RSV_4^7 = (TO,0),\ RSV_6^7 = (VH,-.13),$$
$$RSV_7^7 = (VH,-.13)\}$$

being $RSV_j^i = g_{2t}(d_j, t_i, (c_i,0))$, and where, for example, the value RSV_2^7 is calculated by means of g_{2t} as follows:

$$RSV_2^7 = g_{2t}(d_2, t_7, (H,0)) =$$
$$\Delta\left(\frac{8 \cdot (7.2-5)}{2 \cdot (8-5)} + \frac{8}{2}\right) = \Delta(6.93) = (s_7 = EH, -.07).$$

Evaluation of subexpressions.

The query q has two subexpressions, $q_1 = (t_5,VH) \vee (t_7,H)$ and $q_2 = (t_6,L) \vee (t_7,H)$. Each subexpression is

in disjunctive form, and thus, we must use an operator ϕ_{2t} with orness(W) > 0.5 (for example, with W = [0.7, 0.3]) to process them. The results that we obtain are the following:

- q_1= (t_5,VH)$\vee$ (t_7,H):

$$\{ RSV_1^1 = (\text{EH},-.28),\ RSV_2^1 = (\text{VH},-.19),$$
$$RSV_4^1 = (\text{VH},-.4),\ RSV_5^1 = (\text{EL},.49),$$
$$RSV_6^1 = (\text{VH},-.45),\ RSV_7^1 = (\text{VH},-.45)\},$$

- q_2 = (t_6,L)$\vee$ (t_7,H):

$$\{ RSV_1^2 = (\text{EH},-.25),\ RSV_2^2 = (\text{H},.24),$$
$$RSV_4^2 = (\text{EH},-.44),\ RSV_6^2 = (\text{M},.13),$$
$$RSV_7^2 = (\text{EH},.25)\},$$

being RSV_j^i the evaluation result of the subexpression q_i with respect to the document d_j, where, for example, the RSV_2^2 is calculated by means of the 2-tuple linguistic OWA operator ϕ_{2t} as follows :

$$RSV_2^2 = \phi_{2t}(RSV_2^6 = (\text{EL},.28),\ RSV_2^7 = (\text{EH},-.07)) =$$
$$\Delta(6.93 \cdot 0.7 + 1.28 \cdot 0.3) = \Delta(5,24) = (H,.24),$$

such that Δ^{-1}(EL,.28) = 1.28 and Δ^{-1}(EH,-.07) = 6.93.

Evaluation of the whole query.

We evaluate the whole query using an operator ϕ_{2t} with orness(W) < 0.5 (e.g. with W = [0.3, 0.7]) given that it is in a conjunctive normal form, obtaining the following relevance results RSV_j for each document d_j:

$$\{\text{RSV}_1 = (\text{EH},-.27),\ \text{RSV}_2 = (\text{H},.41),$$
$$\text{RSV}_4 = (\text{VH},-.11),\ \text{RSV}_5 = (\text{N},.45),$$
$$\text{RSV}_6 = (\text{H},-.44),\ \text{RSV}_7 = (\text{VH},.06)\}.$$

5 Concluding remarks

In this paper we have described a new modelling of a linguistic weighted IRS. The IRS has been tuned using the 2-tuple representation model and the 2-tuple symmetrical matching function defined in [8]. With this modelling, the relevance degrees of the final retrieved documents are improved. In the future, we shall research the impact both of the 2-tuple representation model and the 2-tuple symmetrical matching function in a linguistic multi-weighted IRS.

References

[1] A. Bookstein, Fuzzy request: An approach to weighted Boolean searches, *Journal of the American Society for Information Science* 31 (1980) 240-247.

[2] G. Bordogna, C. Carrara and G. Pasi, Query term weights as constraints in fuzzy information retrieval, *Information Processing & Management* 27 (1991) 15-26.

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

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

[5] D. Buell and D.H. Kraft, Threshold values and boolean retrieval systems, *Information Processing & Management* 17 (1981) 127-136.

[6] D. Buell and D.H. Kraft, A model for a weighted retrieval system, *Journal of the American Society for Information Science* 32 (1981) 211-216.

[7] C.S. Cater and D.H. Kraft, A generalization and clarification of the Waller-Kraft wish list, *Information Processing & Management* 25 (1989) 15-25.

[8] E. Herrera-Viedma, Modelling 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.

[9] E. Herrera-Viedma, An information retrieval system with ordinal linguistic weighted queries based on two weighting elements. *Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems* 9 (2001) 77-88.

[10] E. Herrera-Viedma, O. Cordón, M. Luque, A.G. Lopez, A.M. Muñoz, A model of fuzzy linguistic IRS based on multi-granular linguistic information, *International Journal of Approximate Reasoning* 34 (2003) 221-239.

[11] E. Herrera-Viedma, A.G. López-Herrera, C. Porcel, Tuning the Matching Function for a Threshold Weighting Semantics in a Linguistic, *International Journal of Intelligence Systems,* in press.

[12] 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.

[13] 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.

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

[15] D.H. Kraft, G. Bordogna and G. Pasi, An extended fuzzy linguistic approach to generalize Boolean information retrieval, *Information Sciences* 2 (1994) 119-134.

[16] S. Miyamoto, Fuzzy Sets in Information Retrieval and Cluster Analysis (Kluwer Academic Publishers, 1990).

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

[18] L.A. Zadeh, The concept of a linguistic variable and its applications to approximate reasoning. Part I, *Information Sciences* 8 (1975) 199-249, Part II, *Information Sciences* 8 (1975) 301-357, Part III, *Information Sciences* 9 (1975) 43-80.