Enrique Herrera-Viedma, Oscar Cordón, Juan Carlos Herrera, María Luque
Dept. of Computer Science and A.I. University of Granada. 18071-Granada, Spain

An IRS Based on Multi-Granular Linguistic Information

Abstract: An information retrieval system (IRS) based on fuzzy multi-granular linguistic information is proposed. The system has an evaluation method to process multi-granular linguistic information, in such a way that the inputs to the IRS are represented in a different linguistic domain than the outputs. The system accepts Boolean queries whose terms are weighted by means of the ordinal linguistic values represented by the linguistic variable "Importance" assessed on a label set *S*. The system evaluates the weighted queries according to a threshold semantic and obtains the linguistic retrieval status values (RSV) of documents represented by a linguistic variable "Relevance" expressed in a different label set *S'*. The advantage of this linguistic IRS with respect to others is that the use of the multi-granular linguistic information facilitates and improves the IRS-user interaction

1. Introduction

Information Retrieval (IR) is a research field refered the storage and retrieval of textual information (Korfhage, 1997; Salton, 1989). IR systems (IRSs) carry out two main activities: i) to store documents by computing index term weights and ii) to retrieve documents by matching user queries and documents. An important question in the IRSs is how to facilitate the IRS-user interaction.

The use of *linguistic variables* (Zadeh, 1975) to represent the input and output information in the retrieval process of IRSs improves considerably the IRS-user interaction (Bordogna and Pasi, 1993) (Kraft et al., 1994) (Herrera-Viedma, 1999; 2001). Usually, the most linguistic IRSs assume that users provide their information needs by means of Boolean queries whose terms are weighted by linguistic values represented by the linguistic variable "Importance" assessed on a label set *S*. IRSs evaluate the linguistic weighted queries and provide the linguistic RSVs of documents represented by the linguistic variable "Relevance" assessed on the same label set *S*. The drawback is that the use of the same label set to express the inputs and outputs of linguistic IRSs diminishes the communication capability in the IRS-user interaction. Furthermore, the above linguistic variables represent different concepts, and thus, it seems necessary to use different linguistic expression domains to model them.

In this paper, we present a linguistic IRS that manages multi-granular linguistic information using an *ordinal fuzzy linguistic approach* (Herrera et al., 1996) (Herrera-Viedma, 2001). The weighted Boolean queries and the RSVs of documents are assessed on label sets with different granularity and/or semantics. The query terms are weighted according to a *threshold semantic*. The Boolean operators AND and OR are modeled by means of the linguistic aggregation operator, LOWA operator (Herrera et al., 1996). The LOWA operator is an *and-or operator*, and this property allows us to introduce a soft computing in the evaluation of queries. The retrieved documents are arranged in linguistic relevance classes, which are identified by ordinal linguistic terms.

To do so, the paper is structured as follows. Section 2 is devoted to introduce the ordinal fuzzy linguistic approach, the concept of multi-granular linguistic information and the LOWA

operator. Then, the IRS based on multi-granular linguistic information is presented in Section 3. Finally, several conclusions are pointed out in Section 4.

2. The Ordinal Fuzzy Linguistic Approach

The *ordinal fuzzy linguistic approach* is an approximate technique appropriate to deal with qualitative aspects of problems (Herrera et al., 1996). An ordinal fuzzy linguistic approach is defined by considering a finite and totally ordered label set $S = \{s_i, i \in H = 0, \ldots, T\}$ in the usual sense and with odd cardinality (7 or 9 labels). The mid term representing an assessment of "approximately 0.5" and the rest of the terms being placed symmetrically around it. The semantic of the linguistic term set is established from the ordered structure of the term 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 linear trapezoidal membership function represented by the 4-tuple (a_i, b_i, α_i β_i) (the first two parameters indicate the interval in which the membership value is 1.0; the third and fourth parameters indicate the left and right widths of the distribution). Furthermore, we require the following properties:

1. – *The set is ordered* : $s_i \geq s_j$ *if* $i \geq j$.
2. – *There is the negation operator* : $Neg(s_i) = s_j$, *with* $j = T - i$.
3. – *Maximization operator* : $MAX(s_i, s_j) = s_i$ *if* $s_i \geq s_j$.
4. – *Minimization operator* : $MIN(s_i, s_j) = s_i$ *if* $s_j \leq s_j$.

2.1. On Multi-Granular Linguistic Information

In any linguistic approach, an important parameter to determine is the granularity of uncertainty, i.e., the cardinality of the linguistic term set *S* used to express the information. The cardinality of *S* must be small enough so as not to impose useless precision on the users, and it must be rich enough in order to allow a discrimination of the assessments in a limited number of degrees.

On the other hand, according to the uncertainty degree that a user qualifying a phenomenon has on it, the linguistic term set chosen to provide his knowledge will have more or less terms. When different users have different uncertainty degrees on the phenomenon, then several linguistic term sets with a different granularity of uncertainty are necessary. Then, we need tools of management of multi-granular linguistic information to model these situations. Different proposals can be found in (Delgado et al., 1998) (Herrera et al., 2000).

In (Delgado et al., 1998) we characterize some transformation functions between the linguistic and numerical expression domains using the concept of the *characteristic values* associated to a label.

Let us consider that for each label s_i we know a set of characteristic values, $CV_i = \{ C_i^1, C_i^2, \ldots, C_i^z\}$, which are crisp values that summarize the information given by s_i, i.e., they support its meaning. We shall assume that $C_i^j \in Supp(s_i) = \{r \in \Re \mid \mu_{si}(r) > 0\}$. Without loss of generality, we can define a set of functions $CF = \{ f_j, j=1,\ldots,z\}$, in such a way that each function f_j associates a characteristic value to each label s_i, i.e., $f_j : F(\Re) \rightarrow \Re$, $f_j(s_i) = C_i^j$, being $F(\Re)$ the set of fuzzy numbers defined on $\Re$ that we can use to characterize the semantic of the labels. Some examples of this function type are:

- *The deffuzification method of gravity center* (Cordón et al., 1997): *:* $f_1(s_i) = [(b_i + \beta_i)^2 + (b_i)^2 - (a_i)^2 - (a_i - \alpha_i)^2 + (b_i + \beta_i)b_i - a_i(a_i - \alpha_i)] \backslash 3(2b_i + \beta_i - 2a_i + \alpha_i)$, and $f_1(s_i) = a_i - \alpha_i$ if $\beta_i = \alpha_i = 0$.
- *The value* (Delgado et al. 1998b)*:* $f_2(s_i) = (2a_i + 2b_i + \alpha_i + \beta_i) \backslash 6$.
- *The maximum value:* $f_3(s_i) = \max\{v \mid \mu_{si}(v) = \text{Sup}\{\mu_{si}(t), \forall t\}\}$.

Definition 1. The linguistic-numerical transformation function, Λ^N *, for any label* s_i *is defined according to the following expression:* $\Lambda^N : S \rightarrow [0,1]$, $\Lambda^N(s_i) = g(f_1(s_i), f_2(s_i),...,f_z(s_i))$*, being g any aggregation operator verifying:* $\min\{v_1, v_2,..., v_z\} \leq g(v_1, v_2,..., v_z) \leq \max\{v_1, v_2,..., v_z\}$.

Therefore Λ^N obtains the real value of a label by means of the aggregation of its respective characteristic values. An example of *g* can be the mean function.

Definition 2. The numerical-linguistic transformation function, Λ^L *, for any numerical value* $r \in [0,1]$ *is defined according to the following expression:* $\Lambda^L : [0,1] \rightarrow S$, $\Lambda^L(r) = s_i$*, being* s_i *a label verifying:* $h(r,s_i) = \min\{h(r, s_p) \mid \forall s_p \in S,$ *with*

$$h(r, s_t) = \begin{cases} z & \text{otherwise} \\ \sum_{j=1}^{z} (r - f_j(s_t))^2 & \text{if} \quad r \in Supp(s_t) \end{cases}$$

In this paper, we use the above transformation functions to define a tool for processing multi-granular linguistic information in the retrieval process of IRS.

2.2. The LOWA Operator

The *Linguistic Ordered Weighted Averaging* (LOWA) is an aggregation operator of ordinal linguistic values based on symbolic computation (Herrera et al., 1996). It acts by direct computation on the labels only taking into account the order of linguistic assessments without considering the associated membership functions.

Definition 3. Let $A = \{ a_1, \ldots, a_m \}$ *be a set of labels to be aggregated, then the LOWA operator,* Φ*, is defined as* $\Phi(a_1, \ldots, a_m) = W \cdot B^T = C^m\{ w_k, b_k, k = 1, \ldots, m \} = w_1 \Theta b_1 \oplus (1 - w_1) \Theta C^{m-1}\{\beta_h, b_h, h = 2, \ldots, m \}$*, where* $W = [w_1, \ldots, w_m]$*, 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* $B = \{ b_1, \ldots, b_m \}$ *is a vector associated to A, such that,* $B = \sigma(A) = \{ a_{\sigma(1)}, \ldots, a_{\sigma(m)} \}$*, where,* $a_{\sigma(j)} \leq a_{\sigma(i)}$ $\forall i \leq j$*, with* σ *being a permutation over the set of labels A.* C^m *is the convex combination operator of m labels and if m=2, then it is defined as* $C^2\{ w_i, b_i, i = 1, 2 \} = w_1 \Theta s_j \oplus (1 - w_1) \Theta s_i = s_k$*, such that* $k = \min \{ T, i + \text{round}(w_1 \cdot (j - i)) \} s_j$, $s_i \in S$, $(j \geq i)$*, being "round" the usual round operation, and* $b_1 = s_j$, $b_2 = s_i$*. If* $w_j = 1$ *and* $w_i = 0$ *with* $i \neq j \forall i$*, then* $C^m\{ w_i, b_i, i = 1, \ldots, m \} = b_j$.

The LOWA operator is an *"or-and"* operator (Herrera et al., 1996). This property allows that the LOWA operator carries out a soft computing in the modelling of MAX and MIN linguistic operators. We use this good characteristic in our linguistic IRS to evaluate the Boolean queries. In order to classify OWA operators in regard to their localisation between

and and *or*, Yager (Yager, 1988) introduced a measure of *orness*, associated with any vector W as follows

$$orness\,(W) = \frac{1}{m-1}\sum_{k=1}^{m}(m-k)w_{k.}$$

Fixed a W, then the nearer an OWA operator is to an *or*, the closer its orness measure is to one; while the nearer it is to an *and*, the closer is to zero. Generally, an OWA operator with much on nonzero weights near the top will be an *orlike* operator (*orness≥ 0.5)*, and when much of the weights are nonzero near the bottom, the OWA operator will be an *andlike*.

3. The IRS Based on Multi-Granular Linguistic Information

In this section we present an IRS that accepts linguistic weighted Boolean queries, supports multi-granular linguistic information and models the Boolean operators in a flexible way.

We assume that the documents $D=\{d_1,...,d_m\}$ are represented by means of index terms $T=\{t_1,...,t_n\}$. Each term has associated an *index term weight* F which describes the subject content of the documents. $F : DxT \rightarrow [0,1]$ is a numerical indexing function that maps a given document d_j and a given index term t_i to a numeric weight between 0 and 1. $F(d_j, t_i)$ is a numerical weight that represents the degree of significance of d_j in t_i .$F(d_j, t_i)= 0$ implies that the document d_j is not at all about the concept(s) represented by index term t_i, $F(d_j, t_i) =1$ implies that the document d_j is perfectly represented by the concept(s) indicated by t_i, and $F(d_j, t_i) \in (0,1)$ represents the different intermediate significance degrees.

3.1. The Linguistic Weighted Boolean Queries

In this IRS each query is expressed as a combination of the weighted index terms which are connected by the logical operators AND ($\wedge$), OR ($\vee$), and NOT ($\neg$) and weighted with ordinal linguistic terms represented by the linguistic variable "Importance" assessed on a label set *S*. Thus, as was done in (Herrera-Viedma, 2001), we assume a set of ordinal linguistic terms *S* to express the linguistic weights.

In this context, a query is any legitimate Boolean expression whose atomic components (atoms) are 2-tuplas $<t_i, c_i>$ belonging to the set, TxS; $t_i \in$ T (set of index terms), c_i is a label of the linguistic variable "Importance", modelling a threshold semantic. Therefore, the set Q of the legitimate linguistic weighted Boolean queries is defined by the following syntactic rules:

1. $\forall\ q =< t_i, c_i > \in TxS \rightarrow q \in Q$.
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 linguistic weighted Boolean queries $q \in Q$ are only those obtained by applying rules 1-4.

3.2.Evaluation Procedure of User Queries

In this subsection, we present how IRS evaluates a user query in a multi-granular linguistic framework, that is, assuming that the values RSV assigned to the documents are represented

by means of the linguistic variable "Relevance" which is assessed on a label set $S' \neq S$. To define the evaluation procedure, previously we have to establish the semantics associated to the weights of user queries.

Particularly, we assume that the weights of query terms are associated to *a symmetrical threshold semantics* (Herrera-Viedma, 2001). This semantics considers that a user can search for documents with a minimally acceptable presence of one term in their representations as in or documents with a maximally acceptable absence of one term in their representations. Then, when a user asks for documents in which the concept(s) represented by a term t_i is (are) with the value *High Importance*, the user would not reject a document with a F value greater than *High*; on the contrary, when a user asks for documents in which the concept(s) represented by a term t_i is (are) with the value *Low Importance*, the user would not reject a document with a F value less than *Low*. In practice, given a request $<t_i, c_i>$, this means that the linguistic query weights that imply the presence of a term in a document $c_i \geq s_{T/2}$ (e.g. *High, Very High*,) it must be treated differently to the linguistic query weights that imply the absence of one term in a document $c_i \geq s_{T/2}$ (e.g. *Low, Very Low*). Then, if $c_i \geq s_{T/2}$ the request $<t_i, w_i>$, is synonymous with the request $<t_i$, at least $c_i>$, which expresses the fact that the desired documents are those having F values as high as possible; and if $c_i < s_{T/2}$ is synonymous with the request $<t_i$, at most $c_i>$, which expresses the fact that the desired documents are those having F values as low as possible.

Then, the evaluation procedure evaluates a linguistic weighted Boolean query in three steps:

1.- Making uniform the information using the transformation functions given in Subsection 2.1. This implies that the numerical index term weights of the documents and the linguistic weights of queries must be expressed in the domain of the linguistic variable "Relevance".

2.- The documents are evaluated according to their relevance only to atoms of the query applying the symmetrical threshold semantic.

3.- 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.

Then, the evaluation procedure is modelled by a linguistic matching function $E^*: Q \times D \rightarrow S'$ that, for a given $q \in Q$ yields for each $d_j \in D$ an ordinal linguistic value $RSV_j = E^*(q,d_j) \in S'$. E^* is defined recursively applying the following rules:

1.- $E^*(q,d_j) = g^1(\Lambda^L(F(d_j,t_i)), \Lambda^L(\Lambda^N(c_i))), \forall\, q=<t_i,c_i>, \forall_j$, $\Lambda^L : [0,1] \rightarrow S'$, $\Lambda^N : S \rightarrow [0,1]$, and $g^1\ S' \times S' \rightarrow S'$ is the linguistic matching function that models the symmetrical threshold semantics (Herrera-Viedma, 2001):

$g^1(s_a, s_b)=$			
s_0	if $s_b \geq s_{T/2}$	$\wedge$	$s_a = s_0$
s_{i1}	if $s_b \geq s_{T/2}$	$\wedge$	$s_0 < s_a < s_b$
s_{i2}	if $s_b \geq s_{T/2}$	$\wedge$	$s_b \leq s_a < s_T$
s_T	if $s_b \geq s_{T/2}$	$\wedge$	$s_a = s_T$
s_T	if $s_b \geq s_{T/2}$	$\wedge$	$s_a = S_0$
Neg (s_{i1})	if $s_b \geq s_{T/2}$	$\wedge$	$s_0 < s_a < s_b$
Neg (s_{i2})	if $s_b \geq s_{T/2}$	$\wedge$	$s_b < s_a < s_T$
s_0	if $s_b \geq s_{T/2}$	$\wedge$	$s_a = s_T$

$i_1 = \text{Max}\{ 0, \text{round } (b - ((b-a)/K))\}$

$i_2 = \text{Min}\{ T, \text{round } (b + ((b-a)/K))\}$ $\qquad K \in \{ 2,3,4, \ldots . b\}$.

K is a sensitivity parameter defined to control the importance of the closeness between $\Lambda^L(F)$ and w_i^1 in the final result. The greater the value of K, the smaller the importance of the value of distance. $K = 1$ means that the symmetrical threshold semantic is not used.

2.- *On the negated queries, $\neg q$.* We assume that the evaluation procedure can only deal with negated atoms $<\neg t_i, c_i>$. This may be easily achieved applying the De Morgan's laws on any query. Then, we define the evaluation of document d_j for a negated weighted atom $<\neg t_i, c_i>$ from the negation of index term weight $F(d_j, t_i)$: $E^*(q, d_j) = g^1(Neg(\Lambda^L(F(d_j,t_i))), \Lambda^L(c_i))$.

3.- $E^*(\wedge_k^{M\geq 2} q_k, d_j) = \Phi(E^*(q_1,d_j), \ldots E^*(q_M, d_j))$, using a weighting vector W in such a way that $orness(W) < 0.5$.

4.- $E^*(\vee_k^{M\geq 2} q_k, d_j) = \Phi(E^*(q_1,d_j), \ldots E^*(q_M, d_j))$, using a weighting vector W in such a way that $orness(W) \geq 0.5$.

4. Conclusions

We have presented a linguistic IRS that supports the use of different label sets to express system inputs (user queries) and outputs (RSVs). In such a way, we have improved the IRS-user interaction.

In the future, we shall study how to improve the performance of IRSs based on multi-granular linguistic information by means of other different tools of processing of information.

References

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

Cordón, O., Herrera, F. & Peregrín, A. (1997). Applicability of the fuzzy operators in the design of fuzzy logic controllers. *Fuzzy Sets and Systems*, 86: 15-41.

Delgado, M., Herrera, F., Herrera-Viedma, E. & Martínez, L. (1998). Combining numerical and linguistic information in group decision making. *Information Sciences*, 7: 177-194.

Delgado, M., Vila, M.A. & Voxman, W. (1998b). On a canonical representation of fuzzy numbers. *Fuzzy Sets and Systems*, 93: 125-135.

Herrera, F., Herrera-Viedma, E. & Martínez, L. (2000). A fusion approach for managing multi-granularity linguistic term sets in decision making. *Fuzzy Sets and Systems*, 114: 43-58.

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

Herrera-Viedma, E. (1999). Modelling the query subsystem of an information retrieval system using linguistic variables, Proc. Fourth ISKO Conference (EOCONSID'99), Granada, Spain, pp. 157–162.

Herrera-Viedma, E. (2001). 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): 460-475.

Korfhage, R. R. (1997). *Information Storage and Retrieval*. New York: Wiley Computer Publishing.

Kraft, D.H., Bordogna, G. & Pasi, G.(1994). An extended fuzzy linguistic approach to generalize boolean information retrieval, *Information Sciences*, 2: 119–134.

Salton, G. (1989). *Automatic Text Processing - The Transformation, Analysis and Retrieval of Information by Computer*. Addison Wesley Publishing Company.

Yager, R.R. (1988). On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Trans. on Systems Man and Cybernetics*, 18(1): 183-190.

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