A Tool to Weigh the Connectives of Queries in a Linguistic Information Retrieval System

E. Herrera-Viedma, S. Alonso
Dept. Computer Sciences & A.I.,
University of Granada, Spain.
viedma,salonso@decsai.ugr.es

A.G. López-Herrera
Dept. Computer Sciences,
University of Jaén, Spain.
aglopez@ujaen.es

C. Porcel
Dept. Computer Sciences & A.N.,
University of Córdoba, Spain
carlos.porcel@uco.es

Abstract

An ordinal fuzzy linguistic Information Retrieval System (IRS) based on a multi-level weighting scheme to represent the user queries, in a more flexible way, is proposed. The IRS accepts Boolean queries that can be weighted simultaneously by means of ordinal linguistic values in two weighting levels: level of terms and level of connectives. In level of terms, the weights are associated to a threshold semantics, and in the level of connectives they are associated to a control semantics acting as modifiers of the action of the Boolean classical connectives AND and OR in the retrieval process. A new family of parameterized soft computing operators, called S-LOWA operators, is introduced for modelling that control semantics in the action of the connectives AND and OR.

Keywords: Information Retrieval, Weighted Queries, Linguistic Modelling.

1 Introduction

Information Retrieval (IR) may be defined, as the problem of the selection of documentary information from storage in response to search questions provided by a user, which are expressed by a query [1, 14]. Information Retrieval Systems (IRSs) deal with documentary bases containing textual, pictorial or vocal information, organized in documents, and process user queries trying to allow the user to access to relevant information in an appropriate time interval. IRSs present three components to carry out this activity [10]: i) *a database*: to store the documents and the index terms, ii) *a query subsystem*: to formulate the user queries, and iii) *an evaluation subsystem*: to obtain the Retrieval Status Value (RSV) 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. By attaching weights in a query, a user can increase his/her expressiveness and provide a more precise description of his/her desired documents. Fuzzy Set Theory provides a *soft computing methodology* for handling uncertain information and a good mathematical basis, which may be used to model and process the weights in the queries. Many authors have proposed fuzzy weighted IRS models assuming numeric weights [2, 3, 6, 7]. However, it seems more natural to characterize the contents of the desired documents by explicitly associating a linguistic weight to elements in a query, such as "important" or "very important", instead of a numerical value. So, some fuzzy linguistic IRS models [4, 5, 10, 11, 12] have been proposed using a *fuzzy linguistic approach* [19, 20, 21] to model the query weights and RSVs, being useful the called *ordinal fuzzy linguistic approach* [9]. As it is shown in [10], this approach allows us to

reduce the complexity of the design of IRSs.

In order to formalize fuzzy weighted querying, we have to agree upon the query elements that a user can weigh and some aspects of the semantics associated to the query weights as well. Most of the existing IRSs use Boolean queries [1, 14]. In this context, each user query is expressed as a combination of the index terms which are connected by the logical connectives AND ($\wedge$), OR ($\vee$), and NOT ($\neg$). Thereby, the retrieval process can be controlled from four different weighting levels [10, 12]: i) *level of individual terms*, ii) *level of sub-expressions*, which are associations of terms related by logical connectives, iii) *level of the whole query,* which is the biggest sub-expression, and iv) *level of logical connectives*. The first three levels are the most often applied by users. Usually, in these weighting levels weights have been interpreted using any of the following four different semantics [3, 10, 12]: i) as a measure of the importance of a specific element in representing the query, or ii) as a threshold to aid in matching a specific document to the query, or iii) as a description of an ideal or perfect document, or iv) as a limit on the amount of documents to be retrieved for a specific element. The weighting level of logical connectives has not been studied very much. However, its use can enable users to represent their requirements better. For example, a connective weight can be an expression of a desired interrelationship between the specified terms in the query, and as such it can be seen as a user parameter that controls the action of the logical connectives in the evaluation of the relevance of documents from query terms.

The main aim of the paper is to present a linguistic IRS based on a multi-level weighted query subsystem that allows users: i) to set the qualitative aspects of the desired documents by mean of a threshold semantics in the level of the terms, and ii) to introduce a control semantics, in the level of connectives, to model the behaviour of the logical connectives in a more flexible way. We introduce a family of parameterized soft computing operators, called S-LOWA operators, which allows us to model the control semantics of the connectives weights.

The paper is set out as follows. The ordinal fuzzy linguistic approach together with the S-LOWA operators are presented in Section 2. The fuzzy weighted linguistic IRS is defined in Section 3. Finally, Section 4 includes our conclusions.

2 The Ordinal Fuzzy Linguistic Approach

The *ordinal fuzzy linguistic approach* is a fuzzy approximate technique appropriate to deal with qualitative aspects of problems [10]. It models linguistic information by means of ordinal linguistic labels supported by a *linguistic variable* [19, 20, 21]. A linguistic variable is defined by means of a syntactic rule and a semantic rule. In an ordinal fuzzy linguistic approach the syntactic rule is defined by considering a finite and totally ordered label set $\mathcal{S} = \{s_i\}, i \in \{0, \ldots, \mathcal{G}\}$ in the usual sense, i.e., $s_i \geq s_j$ if $i \geq j$, and with odd cardinality (such as 7 or 9 labels), where the mid term represents an assessment of "approximately 0.5", and the rest of the terms being placed symmetrically around it. The semantics of the linguistic term set is established from the ordered structure of the term set by considering that each linguistic term for the pair $(s_i, s_{\mathcal{G}-i})$ is equally informative. In any linguistic approach we need operators of management of linguistic information, such as: i) a *minimization operator*, $MIN(s_a, s_b) = s_a \; if \; a <= b$, ii) a *maximization operator* $MAX(s_a, s_b) = s_a \; if \; a >= b$, iii) a *negation operator* $NEG(s_i) = s_j \mid j = \mathcal{G} - i$, and iv) some aggregation operators, for example the LOWA operator [9].

2.1 The LOWA Operator

Definition 1. *Let $A = \{a_1, \ldots, a_m\}$ be a set of labels to be aggregated, then the LOWA operator, ϕ, is defined as $\phi(a_1, \ldots, a_m) = W \cdot B^T = \mathcal{C}^m\{w_k, b_k, k = 1, \ldots, m\} = w_1 \odot b_1 \oplus (1 - w_1) \odot \mathcal{C}^{m-1}\{\beta_h, b_h, h = 2, \ldots, m\}$ where $W = [w_1, \ldots, w_m]$, is a weighting vector, such that, $w_i \in [0, 1]$ and $\Sigma_i w_i = 1$.*

$\beta_h = w_h/\Sigma_2^m w_k, h = 2,\ldots,m$, *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.* $\mathcal{C}^m$ *is the convex combination operator of m labels and if m=2, then it is defined as* $\mathcal{C}^2\{w_i, b_i, i = 1,2\} = w_1 \odot s_j \oplus (1 - w_1) \odot s_i = s_k$, *such that* $k = min\{\mathcal{G}, i + round(w_1 \cdot (j - i))\}$ s_j, $s_i \in S$, $(j \geq i)$ *where "round" is 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 the convex combination is defined as:* $\mathcal{C}^m\{w_i, b_i, i = 1,\ldots,m\} = b_j$.

The behavior of the LOWA operator can be controlled by means of the weighting vector W. For example,

$\phi(a_1,\ldots,a_m) = MAX_i(a_i)$ if $W^* = [1,\ldots,0]$,

$\phi(a_1,\ldots,a_m) = MIN_i(a_i)$ if $W_* = [0,\ldots,1]$,

$\phi(a_1,\ldots,a_m) = Ave(a_i)$ if $W_A = [\frac{1}{m},\ldots,\frac{1}{m}]$.

In order to classify OWA operators with respect to their location between *and* and *or*, Yager [17] introduced a measure to characterize the type of aggregation for a particular weighting vector W. This measure, called *orness measure* of the aggregation, is defined as

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

This measure, which lies in the unit interval, characterizes the degree to which the aggregation is like an *or* (MAX) operation. It can be easily shown that $orness(W^*) = 1$, $orness(W_*) = 0$, and $orness(W_A) = .5$. Note that the nearer W is to an *or*, the closer its measure is to one; while the nearer it is to an *and*, the closer is to zero. Therefore, as we move weight up the vector we increase the $orness(W)$, while moving weight down causes us to decrease $orness(W)$. We can easily see that the dual operator of an OWA operator defined with weighting vector $W^\wedge = [w_i^\wedge = w_{m-i+1}]$ satisfies that

$$orness(W) = 1 - orness(W^\wedge),$$

and therefore, if an OWA operator is *orlike* then its dual is *andlike*. The *andness measure* can be defined from the orness measure as [17] $andness(W) = 1 - orness(W)$.

2.2 The S-LOWA Operators

In our linguistic weighted IRS we need to aggregate ordinal fuzzy linguistic information and at the same time to interpret the connective weights. To do so, we introduce a new family of operators based on the LOWA operators [9], called S-LOWA operators.

The problem of the OWA operators is the determination of the weighting vector. A number of approaches have been suggested for obtaining the weights [16, 17]. Some of them allow the participation of users in the procedure for calculating the weights. In such cases, the behaviour of OWA operator may be guided or controlled by the user's preferences. One of these procedures consists of generating the weights from parameters provided by the users. In [18] were presented two parameterized OWA operators, denoted *S-OWA operators*, which can learn weighting vector from the *orness* and *andness* expressed by a user, respectively. The first operator is an *orlike S-OWA operator* with weighting vector W^{SO} defined as

$$w_1 = \frac{2 - 2\cdot\alpha}{m} + 2\cdot\alpha - 1, \alpha \in [0.5, 1],$$

$$w_i = \frac{2 - 2\cdot\alpha}{m}, \text{for } i = 2,\ldots,m,$$

with $\alpha = orness(W^{SO})$. The second one is an *andlike S-OWA operator* with weighting vector W^{SA} defined as

$$w_m = \frac{2 - 2\cdot\alpha}{m} + 2\cdot\alpha - 1, \alpha \in [0.5, 1],$$

$$w_i = \frac{2 - 2\cdot\alpha}{m}, \text{for } i = 1,\ldots,m-1,$$

with $\alpha = andness(W^{SA})$. When $\alpha = 0.5$ both OWA operators reduce to the arithmetic mean operator.

Then, in the evaluation of the user queries we shall use an *andlike S-LOWA operator* (ϕ^{SA}) and an *orlike S-LOWA operator* (ϕ^{SO}) to model the soft computing of the query logical connectives *AND* and *OR*, respectively.

3 A Weighted Linguistic IRS

In this Section, we present a weighted linguistic IRS model using the above ordinal fuzzy linguistic approach. This IRS presents a multi-level weighting scheme for formulating the user queries. In particular, it allows users to weigh the query terms and connectives. With a such scheme users can control better the retrieval of their desired documents.

3.1 The Documentary Database

$\mathcal{D} = \{d_1, \ldots, d_m\}$ is a finite set of documents or records. Each document is represented by means of a finite set of index terms $\mathcal{T} = \{t_1, \ldots, t_l\}$. The index terms describe the subject content of each document by means of a numeric indexing function $\mathcal{F} : \mathcal{D} \times \mathcal{T} \rightarrow [0, 1]$. Then, each document d_j is represented as a fuzzy subset of $\mathcal{T}$ characterized by the membership function $\mathcal{F}$, $d_j = \sum_{i=1}^{l} \mathcal{F}(d_j, t_i)/t_i$.

3.2 The Query Subsystem

The query subsystem accepts weighted Boolean queries whose query weights are ordinal linguistic values. By assigning weights in queries, users specify restrictions on the documents that the IRS has to satisfy in the retrieval activity. We observe that in a typical Boolean query there are four possible weighting levels [10, 12]: the level of terms, the level of sub-expressions, the level of whole query and the level of the Boolean connectives AND and OR. Most defined IRSs support mainly the first three weighting levels, although not simultaneously. However, it is obvious that the retrieval activity strongly depends on the operators used to model the action of connectives. Therefore, the use of the fourth level would allow users to control the action of operators and guide better the retrieval of their desired documents.

We assume that users can simultaneously use two weighting levels, terms and connectives, to express their desired documents. Accordingly, the set of the legitimate queries $\mathcal{Q}$ is defined by the following syntactic rules:

1.- $\forall q =< t_i | c^1 > \rightarrow q \in \mathcal{Q}$, where $t_i \in \mathcal{T}$ and $c^1 \in \mathcal{S}$ is the ordinal linguistic weight assigned by a user in the weighting level of index terms. This rule defines simple queries.

2.- $\forall q =< \wedge_{k=1}^{n \geq 2} q_k, c^2 >, q_k \in \mathcal{Q} \rightarrow q \in \mathcal{Q}$, where $c^2 \in \mathcal{S}$ is the ordinal linguistic weights assigned by a user in the weighting level of connectives to combine terms in the subexpressions. This rule defines the queries expressed by conjunctive queries AND.

3.- $\forall q =< \vee_{k=1}^{n \geq 2} q_k, c^2 >, q_k \in \mathcal{Q} \rightarrow q \in \mathcal{Q}$. This rule defines the queries expressed by disjunctive queries OR.

4.- $\forall q \rightarrow \neg q \in \mathcal{Q}$. This rule defines negated queries.

5.- All legitimate queries are only those obtained by applying rules 1-4, inclusive.

We should point out that all ordinal linguistic weights used in a query are terms of the linguistic variable *Importance*, but modeling different semantics or interpretations depending on the weighting level.

To sum up, we propose a query subsystem with a multi-level weighted query language which manages two possible weighting levels. Then, in the formulation of any query the users can assign two kinds of weights: 1) weights on query terms which are associated to a *threshold semantics*, and 2) weights on query connectives which are associated to a *control semantics*. By associating threshold weights [6, 7, 13] with terms in a query, the user is asking to see all the documents sufficiently related to the topics represented by such terms. The weights in the connectives can act as modifiers of the action of classical connectives AND and OR. By assigning weights in the connectives of a query the users can carry out a soft control on the retrieval of system in order to guide its action towards their desired documents. The control semantics defines weights of connectives AND and OR as *andness* and *orness* measures that control the restrictive and inclusive behaviour of the connectives AND and OR in the computation of RSVs, respectively.

3.3 The Evaluation Subsystem

The evaluation subsystem is implemented by the matching or evaluation function $\mathcal{E}$, which assesses the relationship between $\mathcal{Q}$ and $\mathcal{D}$ by means of linguistic RSVs taken from the linguistic variable *"Relevance"*. Therefore, the goal of $\mathcal{E}$ consists of evaluating documents in terms of their relevance to a multi-level weighted query according to two weighting levels. We define $\mathcal{E}$ by means of a constructive bottom-up evaluation process that satisfies the *criterion of separability* [8, 15] at the same time as supporting all the weighting semantics.

The evaluation function $\mathcal{E}$ acts in two steps: 1) firstly, the documents are evaluated according to their relevance only to atoms of the query. In this step, a partial RSV is assigned to each document with respect to each atom, and 2) the documents are evaluated according to their relevance to Boolean combinations of atomic components (their partial RSVs), and so on, working in a bottom-up method until the whole query is processed. In this step, a total RSV is assigned to each document with respect to the whole query. Therefore, a set of linguistic terms $\mathcal{S}$ is used to represent the relevance values.

Then, given any query $q \in \mathcal{Q}$, we define $\mathcal{E} : \mathcal{D} \times \mathcal{Q} \rightarrow \mathcal{S}$ according to the following four evaluation rules:

1. If $q = \langle t_i, c^1 \rangle$ then

$$\mathcal{E}(d_j, q) = g(d_j, t_i, c^1),$$

where $g : \mathcal{D} \times \mathcal{T} \times \mathcal{S} \rightarrow \mathcal{S}$ is the linguistic matching function to model the threshold semantics defined to the following expression:

$$g(d_j, t_i, c^1) = \begin{cases} s_b & \text{if } s_a \geq c^1 \\ s_c & otherwise. \end{cases}$$

where $s_a = Label(\mathcal{F}(d_j, t_i))$, $Label : [0,1] \rightarrow \mathcal{S}$ is a function that assigns a label in $\mathcal{S}$ to a numeric value $r \in [0,1]$ according to the expression: $Label(r) = s_i$ with $i = round(\mathcal{G} \cdot r)$, being $round(\cdot)$ the usual "round" operator; $b = min(\mathcal{G}, a + round(2 \cdot \frac{\mathcal{G}-a}{\mathcal{G}}))$; and $c = max(0, a - round(2 \cdot \frac{\mathcal{G}-a}{\mathcal{G}}))$.

2. If $q =< \wedge_{k=1}^{n \geq 2} q_k, c^2 >, q_k \in \mathcal{Q}$, then

$$\mathcal{E}(d_j, q) = \phi^{SA}(RSV_{1j}, \ldots, RSV_{kj}),$$

with $RSV_{kj} = \mathcal{E}(d_j, q_k) \forall k$.

3. If $q =< \vee_{k=1}^{n \geq 2} q_k, c^2 >, q_k \in \mathcal{Q}$, then

$$\mathcal{E}(d_j, q) = \phi^{SO}(RSV_{1j}, \ldots, RSV_{kj}),$$

with $RSV_{kj} = \mathcal{E}(d_j, q_k) \forall k$.

4. If q is negated then

$$\mathcal{E}(d_j, \neg q) = NEG(\mathcal{E}(d_j, q)).$$

When the evaluation subsystem finishes, the IRS presents the retrieved documents arranged in linguistic relevance classes in decreasing order of $\mathcal{E}$, in such a way, that the maximal number of classes is limited by the cardinality of the set of labels chosen for the linguistic variable *Relevance*.

4 Conclusions

In this paper, we have presented an ordinal fuzzy linguistic IRS model that accepts multi-level weighted Boolean queries and returns documents arranged in relevance classes labeled with ordinal linguistic values. Its main advantage with respect to other IRSs is that users can specify better the characteristics of documents that they desire by means of two levels of weighting: level of terms and level of connectives. In such a way, users control or guide better the retrieval process of IRS in order to effectively retrieve documents satisfying their concepts of relevance.

Acknowledgements

This work has been supported by the project Proyecto de Excelencia de la Junta de Andalucia SAINFOWEB, Cod. 00602.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*, Addison-Wesley, 1999.

[2] A. Bookstein, *Fuzzy request: An approach to weighted boolean searches*,

Journal of the American Society for Information Science and Technology (1980), no. 31, 240–247.

[3] G. Bordogna, P. Carrara, and G. Pasi, *Query term weights as constraints in fuzzy information retrieval*, Information Processing and Management **27** (1991), no. 1, 15–26.

[4] 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), no. 2, 70–82.

[5] ———, *An ordinal information retrieval model*, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **9** (2001), 63–76.

[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] ———, *Threshold values and boolean retrieval systems*, Information Processing and Management **17** (1981), 127–136.

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

[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] 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** (2001), no. 6, 460–475.

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

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

[13] D.H. Kraft and D.A. Buell, *Fuzzy sets and generalized boolean retrieval systems*, International Journal of Man-Machine Studies **19** (1983), 45–56.

[14] G. Salton and M.J. McGill, *An introduction to modern information retrieval*, McGraw-Hill, 1983.

[15] W.G. Waller and D.H. Kraft, *A mathematical model of a weighted boolean retrieval system*, Information Processing and Management **15** (1979), 235–245.

[16] Z. Xu, *An overview of methods for determining* OWA *weights*, International Journal of Intelligent Systems **20** (2005), 843–865.

[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] R.R. Yager and D.P. Filev, *Parametrized "andlike" and "orlike" owa operators*, International Journal of General Systems **22** (1994), 297–316.

[19] L. A. Zadeh, *The concept of a linguistic variable and its applications to approximate reasoning.* P*art* I, Information Sciences **8** (1975), 199–249.

[20] ———, *The concept of a linguistic variable and its applications to approximate reasoning.* P*art* II, Information Sciences **8** (1975), 301–357.

[21] ———, *The concept of a linguistic variable and its applications to approximate reasoning.* P*art* III, Information Sciences **9** (1975), 43–80.