Modelling the Query Subsystem of an Information Retrieval System Using Linguistic Variables

E. Herrera-Viedma
Dept. of Computer Science and Artificial Intelligence
University of Granada, 18071 - Granada, Spain
e-mail: viedma@decsai.ugr.es

Abstract:

A linguistic model of the query subsystem of an information retrieval systems based on the concept of linguistic variables [12] is presented. Queries are weighted by means of the linguistic weights expressing a semantic of importance. A weighted query evaluation mechanism based on the Linguistic Weighted Disjunction operator and the Linguistic Weighted Conjunction operator [7] is given.

Resumen:

En este trabajo se presenta un modelo lingüístico del módulo de consultas de un sistema de recuperación de información documental. Este modelo se realiza usando el concepto de variable lingüística introducido por Zadeh [12]. Las consultas son ponderadas mediante el uso de pesos lingüísticos expresando una semántica de importancia. Para evaluar las consultas, se define un método de evaluación basado en dos operadores de agregación de información lingüística ponderada, uno para el conectivo AND y otro para OR [7].

Keywords: fuzzy information retrieval, linguistic modelling, weighted queries.
Palabras clave: Variables lingüísticas, consultas ponderadas, recuperación de información difusa.

1. Introduction

The Information Retrieval (IR) involves the development of computer systems for the storage and retrieval of (predominantly) textual information (documents). An IR system is an information system whose aim function is to evaluate user queries for information based on a content-analysis of the archived documents. Basically, a IR system presents three parts:

1. ***A Database****:* which stores the documents and the representation of their information content. It is built using a method for extracting and representing the documents contents.

2. ***A Query Subsystem:*** which allows to the users to formulate their queries and presents the relevant documents retrieved by the system for these queries. To do that, it must have a query language to build the queries and procedures to choice the relevant documents.

3. ***A Evaluation Subsystem****:* which retrieves and evaluates the relevant documents for an user query. This activity is achieved by means of an inference procedure that establishes a relationship between the user request and the documents in the collection to determine whether a document is relevant or not.

The query subsystem supports the user-IR system interaction, and therefore, it should be able to account for 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 provide a more precise description of his or her information needs. Many authors have dealt with this aspect within Fuzzy Set Theory [2,3,4,5,8,10,11]. In most these fuzzy retrieval models [2,3,5,8,10,11], the users use numeric weights (values in [0,1]) in the weighted queries. As a result of the evaluation of a weighted query to each document retrieved is assigned a numeric value, called its Retrieval Status Value (RSV), which indicates the estimated relevance of the document to the user information needs.

However, the use of query languages based on numeric weights forces the user to quantify qualitative concept (such as "importance"), ignoring that many users are not able to provide their information needs precisely in a quantitative form but in a qualitative one. In fact, it seems more natural to characterize the desired document contents by explicitly associating to a term in a query a linguistic descriptor like "important" or "very important", instead of a numerical value. In the same way, the IR system is more friendly if the query subsystem supplies the estimated relevance of the documents in a linguistic form, (e.g., linguistic terms like "relevant", "very relevant", "fairly relevant" can be used) rather than by scores (the RSVs). In [4,9] a *fuzzy linguistic approach* is given for modelling the query subsystem in this sense.

The fuzzy linguistic approach is an approximate technique, which represents qualitative aspects as linguistic values by means of *linguistic variables*, that is, variables whose values are not numbers but words or sentences in a natural or artificial language [12]. Using linguistic variables, each linguistic value is characterized by a "syntactic value" or "label" and a "semantic value" or "meaning". The label is a word or sentence belonging to a linguistic term set, and the meaning is a fuzzy subset in an universe of discourse.

On the other hand, in order to formalize fuzzy weighted quering, the semantic associated to the query weights must be established. Mainly there are three possibilities [3,9]: (i*) importance or relative relevance*, (ii) *thresholds*, and (iii) *description of an ideal or perfect document*.

The main of this paper is to present a new fuzzy linguistic approach to model the query subsystem. We propose to use a non-classical fuzzy linguistic approach [7] to represent the linguistic weights of the user's queries and the linguistic RSVs of the retrieved documents. In such a way, we overcomes the limitations of the classical fuzzy linguistic approach, i.e, we have not to establish explicitly semantic rules neither syntactic rules. The weights are assigned in two distinct levels, query term weight and query clause weight, with the same semantic. We consider that the semantic of the weights is importance and give an RSV evaluation mechanism based on two aggregation operators of weighted linguistic information, the ***Linguistic Weighted Disjunction (LWD)*** operator and the ***Linguistic Weighted Conjunction (LWC)*** operator [7].

The paper is set out as follows. The fuzzy linguistic approach is presented in Section 2. The linguistic modelling of the query subsystem is given in Section 3. Finally, Section 4 draws our conclusions.

2. The Fuzzy Linguistic Approach

The fuzzy linguistic approach is an approximate technique, which represents qualitative aspects as linguistic values by means of linguistic variables, that is, variables whose values are not numbers but words or sentences in a natural or artificial language [12]. Its application is beneficial because it introduces a more flexible framework which allows us a representation of the information in a more direct and adequate way when we are unable to express it with precision. In this way, the burden of quantifying a qualitative concept is eliminated.

The choice of the linguistic term set with its semantic is the first goal to satisfy in any linguistic approach for solving a problem. It consists of establishing the linguistic variable or linguistic expression domain with a view to provide the linguistic performance values.

Definition 1 [12].- A linguistic variable is characterized by a quintuple *(L,H(L),U,G,M)* in which *L* is the name of the variable; *H(L)* (or simply *H*) denotes the term set of *L*, i.e., the set of names of linguistic values of *L*, with each value being a fuzzy variable denoted generically by *X* and ranging across a universe of discourse *U* which is associated with the base variable *u*; *G* is a syntactic rule (which usually takes the form of a grammar) for generating the names of values of *L*; and *M* is a semantic rule for associating its meaning with each *L*, *M(X)*, which is a fuzzy subset of *U*.

From a practical point of view, we can find two possibilities to choose the appropriate linguistic descriptors of the term set and their semantic:

1. The first possibility defines the linguistic term set by means of a context-free grammar, and the semantic of linguistic terms is represented by fuzzy numbers described by membership functions based on parameters and a semantic rule [4,9,12].

2. The second one defines the linguistic term set by means of an ordered structure of linguistic terms, and the semantic of linguistic terms is derived from their own ordered structure which may be either symmetrically distributed on the [0,1] interval or not [7].

In both possibilities, in order to establish the linguistic descriptors of a linguistic variable, an important aspect to analyze is the *granularity of uncertainty*, i.e., the cardinality of the linguistic term set used to express the information. The cardinality of the term set 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. Typical values of cardinality used in the linguistic models are odd ones, such as 7 or 9, with an upper limit of granularity of 11 or no more than 13, where the mid term represents an assessment of "approximately 0.5", and with the rest of the terms being placed symmetrically around it [1]. In the first possibility the granularity of uncertainty is not easily under control, and we can find inadequate cardinalities (very high). However, in the second one, we can control this aspect and supply users with a few but meaningful linguistic descriptors.

In this paper, we assume the second possibility in order to reduce the complexity of defining a grammar and a semantic rule and to control and supervise the granularity of uncertainty. We consider a finite and totally ordered label set in the usual sense and with odd cardinality as in [1]:

$$S = \{s_i, i \in H = 0, \ldots, T\}$$

The mid term representing an assessment of "approximately 0.5" and the rest of the terms are placed symmetrically around it, and assuming that each linguistic term for the pair (s_v, s_{T-v}) is equally informative. The semantic of the labels is given by fuzzy numbers defined on the [0,1] interval, which are described by linear trapezoidal membership functions represented by the 4-tuple:

$$(a_v, b_v, \alpha_v, \beta_v)$$

(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:

$$1. - \textit{The set is ordered} : s_i \geq s_j \quad \textit{if} \quad i \geq j.$$

$$2. - \textit{There is the negation operator} : Neg(s_i) = s_j, \textit{ with } j = T - i.$$

$$3. - \textit{Maximization operator} : MAX(s_i, s_j) = s_i \textit{ if } s_i \geq s_j.$$

$$4. - \textit{Minimizationoperator} : MIN(s_i, s_j) = s_i \textit{ if } s_j \leq s_j.$$

For example, we can use the following set of seven labels with its associated semantic to evaluate the linguistic variables "importance" and "relevance" in our query subsystem:{P = Perfect = (1, 1 ,.16, 0), VH = Very High = (.84, .84, .18 , .16), H = High = (.66, .66 ,.16 ,.18), M = Medium = (.5, .5, .16, .16), L = Low = (.34, .34, .18, .16), VL = Very Low = (.16, .16, .16, .18), N = None = (0, 0, 0, .16)}.

On the other hand, the management of linguistic information requires the use of the adequate aggregation operators of linguistic information. A technique to combine linguistic values given on an ordered set of labels like S is the symbolic approach [7]. It acts by direct computation on labels by taking into account the meaning and features of such linguistic assessments.

In order to evaluate the RSVs, we use the following aggregation operators of linguistic weighted information [7]:

Definition 2.- The aggregation of a set of linguistic weighted opinions, $\{(c_1, a_1), \ldots (c_m, a_m)\}, c_i, a_i \in S,$ according to the *Linguistic Weighted Disjunction (LWD)* operator is defined as:

$$LWD((c_1, a_1), , (c_m, a_m)) = MAX_{i=1,\ldots,m} MIN(c_i, a_i),$$

where a_i shows the weighted opinion, and c_i the importance degree of a_i.

Definition 3.- The aggregation of set of linguistic weighted opinions $\{(c_1,a_1),\ldots(c_m,a_m)\}$, $c_i,a_i \in S$, according to the *Linguistic Weighted Conjunction (LWC)* operator is defined as:

$$LWC((c_1,a_1),,(c_m,a_m)) = MIN_{i=1,\ldots,m} MAX(Neg(c_i),a_i).$$

3. The Linguistic Model of the Query Subsystem

Fuzzy IR systems can be viewed as a formalization of the weighted Boolean IR approach [6]. Weighted Boolean IR systems introduce numeric weights to improve both document representation (index term weights) and query language (query weights). Under this perspective, we define the parts of our IR systems as a natural linguistic extension of the weighted Boolean model. In particular, we present a new linguistic approach to model the query subsystem and the search subsystem.

3.1. Definition of the Database

The database stores the documents and their representations. The document representation is typically based on index terms which are the atomic componets of documents. We assume that the IR system has all mechanisms and data structures necessary to store the documents *D* and the index terms *T* in archives, and also methods to extract the index terms from documents.

We consider that the document representation, *R(di)*, is a fuzzy subset defined in *T*, which is characterized by a membership function $\mu_{R(d_i)} : T \to [0,1]$ i.e.,

$$R(d_i) = \sum_{t_j \in T} \mu_{R(d_i)}(t_j),$$

where $\mu_{R(d_i)}(t_j)$ is a numerical weight that represents the degree of significance of *tj* in *di*.

The quality of the retrieval results strongly depends on the criteria used to automatically compute the index term weights. Different document term weighting schemes can be found in [3,6]. In this paper, we do not focus this aspect and assume any of weighting methods.

3.2. Definition of the Linguistic Query Subsystem

The linguistic query subsystem proposed is an linguistic extension of the weighted Boolean query subsystem. In a weighted Boolean IR system 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 ($\neg$).

We assume that the weights $C=\{ci\}$, $ci \in S$. The weights act in the user queries on single terms and on clauses (terms connected with logical connectives). We adopt a semantic for these linguistic weights that defines them as measures of the importance of each term (clause) with respect to the others existing in the query [2]. With such a semantic, the user requires that the computation of the RSVs is dominated by the more heavily weighted terms and clauses in the query.

We denote by *Q* the set of weighted legitimate queries that can be formulated by means of the query language. *Q* is generated by the following rules:

1.- $\forall q_0 = (t_i,c_i) \in T \times C \Rightarrow q_0 \in Q.$

2.- $\forall (t_1,c_1),(t_2,c_2),\ldots,(t_n,c_n)_{(n\geq 2)} \Rightarrow q_1 = (t_1,c_1) \wedge (t_2,c_2) \wedge \ldots \wedge (t_n,c_n) \in Q.$

3.- $\forall (t_1,c_1),(t_2,c_2),\ldots,(t_n,c_n)_{(n\geq 2)} \Rightarrow q_2 = (t_1,c_1) \vee (t_2,c_2) \vee \ldots \vee (t_n,c_n) \in Q.$

4.- $\forall (q_i,c_i),(q_k,c_k) \in Q \times C, i,k \in \{1,2\} \Rightarrow q_3 = (q_i,c_i) \wedge (q_k,c_k) \in Q.$

5.- $\forall (q_i,c_i),(q_k,c_k) \in Q \times C, i,k \in \{1,2\} \Rightarrow q_4 = (q_i,c_i) \vee (q_k,c_k) \in Q.$

6.- $\forall q \in Q \Rightarrow q_5 = \neg q \in Q.$

7.- All legitimate weighted queries $q \in Q$ are only those obtained by applying rules 1 - 6.

On the other hand, the query subsystem presents the retrieved documents arranged in relevance classes as in [4], but reducing the complexity of the classification process. The maximal number of the classes will be limited by the cardinality of the set of labels chosen. Then, we denote the set of retrieved documents for a query $q \in Q$ as

$$f(q) = \{(d_i, \mu_{f(q)}(d_i)), \forall i)\}, \mu_{f(q)}(d_i) \in S.$$

Therefore, *f(q)* is a subset fuzzy defined in D and characterized by a linguistic membership $\mu f(q)$, such that, $\mu f(q)(di)=RSVi$. This definition of *f(q)* is a formal one, that implies that all documents are presented to the user. However, we propose to show only the more relevant classes of documents.

3.3. Definition of the Evaluation Subsystem

The evaluation subsystem must have a query evaluation mechanism that acts in the retrieval process of relevant documents. This evaluation mechanism assigns the RSVs to the documents. It must be defined according to the semantic adopted for the weights.

The evaluation mechanism is represented by means of a matching function *F* [3,4,6]. We have six query kinds, *(q0,q1,q2,q3,q4,q5),* then, $\forall d\_j \in D, F:QxD \rightarrow S$ is defined as follows:

1.- $F(q_0, d_j)$ is defined by means of a function $label : [0,1] \rightarrow S$ that assigns a label to a numeric value, i.e.,

$$F(q_0, d_j) = label(\mu_{R(d_j)}(t_i)) = Sup_q \{s_q \in S : \mu_{s_q}(\mu_{R(d_j)}(t_i)) = Sup_{v \in H} \{\mu_{s_v}(\mu_{R(d_j)}(t_i))\}\}$$

2.– $F(q_1, d_j) = LWC((label(\mu_{R(d_j)}(t_1)), c_1), (label(\mu_{R(d_j)}(t_2)), c_2), \ldots, (label(\mu_{R(d_j)}(t_n)), c_n))$.

3.– $F(q_2, d_j) = LWD((label(\mu_{R(d_j)}(t_1)), c_1), (label(\mu_{R(d_j)}(t_2)), c_2), \ldots, (label(\mu_{R(d_j)}(t_n)), c_n))$.

4.– $F(q_3, d_j) = LWC((\mu_{f(q_i)}(d_j), c_i), (\mu_{f(q_k)}(d_j), c_k))$.

5.– $F(q_4, d_j) = LWD((\mu_{f(q_i)}(d_j), c_i), (\mu_{f(q_k)}(d_j), c_k))$.

6.– $F(q_5, d_j) = Neg(\mu_{f(q)}(d_j))$.

We should point out that the operators LWC and LWD guarantees that the more important the query terms, the more influential they are in the determination of the RSVs.

4.- Conclusions

We have presented a linguistic model of the query subsystem of a IR system. It incorporates linguistic weights expressing a semantic of importance in two levels: query terms and query clauses. We have used a fuzzy linguistic approach that reduces the complexity of the design process of IR system. We have defined a weighted query evaluation mechanism that overcomes some limitations of classical evaluation mechanisms (see [3,9] problems of the connective AND).

References

1. P.P. Bonissone and K.S. Decker, *Selecting Uncertainty Calculi and Granularity: An Experiment in Trading-off Precision and Complexity*, in: L.H. Kanal and J.F. Lemmer, Eds., Uncertainty in Artificial Intelligence} (North-Holland, 1986) 217-247.

2. A. Bookstein, Fuzzy Request: and Approach to Weighted Boolean Searches, *J. Am. Soc. Information Sci*. **31** (1980) 240-247.

3. G. Bordogna, C. Carrara and G. Pasi, Query Term Weights as Constraints in Fuzzy Information Retrieval, *Information Process. Management* **1** (1991) 15-26.

4. G. Bordogna and G. Pasi, A Fuzzy Linguistic Approach Generalizing Boolean Information Retrieval: A Model and Its Evaluation, *J. Am. Soc. Information Sci*. **44** (1993) 70-82.

5. D. Buell and D.H. Kraft, A Model for a Weighted Retrieval System, *J. Am. Soc. Information Sci*. **32** (1981) 211-216.
6. V Cross, Fuzzy Information Retrieval, *J. of Intelligent Information Systems* **3** (1994) 29-56.

7. F. Herrera and E. Herrera-Viedma, Aggregation Operators for Linguistic Weighted Information, IEEE Transactions on Systems, Man, and Cybernetics} {\bf 27} (1997) 646-656.

8. P.B. Kantor, The Logic of Weighted Queries, *IEEE Transaction on Systems Man and Cybernetics* **11** (1981) 816-821.

9. D.H. Kraft, G. Bordogna and G. Pasi, An Extended Fuzzy Linguistic Approach to Generalize Boolean Information Retrieval, *Information Sciences* **2** (1994) 119-134.

10. G. Salton, E.A. Fox and H. Wu, Extended Boolean Information Retrieval, *Communications of the ACM* **26** (1983) 1022-1036.

11. W.G. Waller and D.H. Kraft, A Mathematical Model of a Weighted Boolean Retrieval System, *Information Process. Management* **15** (1979) 235-245.

12. 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.