PERGAMON

Information Processing and Management 39 (2003) 233–249

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Evaluating the informative quality of documents in SGML format from judgements by means of fuzzy linguistic techniques based on computing with words

Enrique Herrera-Viedma [a,*], Eduardo Peis [b]

[a] *Department of Computer Science and A.I., Library Science Studies School, University of Granada, 18071 Granada, Spain*

[b] *Department of Library Science Studies, Library Science Studies School, University of Granada, 18071 Granada, Spain*

Abstract

Recommender systems evaluate and filter the great amount of information available on the Web to assist people in their search processes. A fuzzy evaluation method of Standard Generalized Markup Language documents based on computing with words is presented. Given a document type definition (DTD), we consider that its elements are not equally informative. This is indicated in the DTD by defining linguistic importance attributes to the more meaningful elements of DTD chosen. Then, the evaluation method generates linguistic recommendations from linguistic evaluation judgements provided by different recommenders on meaningful elements of DTD. To do so, the evaluation method uses two quantifier guided linguistic aggregation operators, the linguistic weighted averaging operator and the linguistic ordered weighted averaging operator, which allow us to obtain recommendations taking into account the fuzzy majority of the recommenders' judgements. Using the fuzzy linguistic modeling the user–system interaction is facilitated and the assistance of system is improved. The method can be easily extended on the Web to evaluate HyperText Markup Language and eXtensible Markup Language documents.

Keywords: SGML; XML; Filtering; Document evaluation; Linguistic modeling; Aggregation

1. Introduction

Internet-based information retrieval is a widely studied and hotly debated topic. Finding relevant, high quality information on the World Wide Web (WWW) is a difficult task. The

* Corresponding author. Tel.: +34-95-824-6251; fax: +34-95-824-3317.
E-mail addresses: viedma@decsai.ugr.es (E. Herrera-Viedma), epeis@ugr.es (E. Peis).

0306-4573/03/$ - see front matter
PII: S0306-4573(02)00049-3

exponential increase in Web sites and Web documents is contributing to that Internet users not being able to find the information they seek in a simple and timely manner. There are many publicly available search engines, but users are not necessarily satisfied with the different formats for inputting queries, speeds of retrieval, presentation formats of the retrieval results, and quality of retrieved information. Therefore, users are in need of tools to help them cope with the mass of content available on the WWW (Kobayashi & Takeda, 2000; Lawrence & Giles, 1998). The development of standard formats for the representation of documents in Web improves substantially the quality of information retrieved by search engines. Furthermore, they help to preserve Web device independence and allow content reuse (Kobayashi & Takeda, 2000; Wium & Saarela, 1999).

The Standard Generalized Markup Language (SGML) is a general metalanguage that can be used to build application specific languages to encode structured documents (Goldfarb, 1990). That is, SGML provides the rules for defining a markup language based on tags. Each instance of SGML includes a description of the document structure called a *document type definition* (DTD). Hence, an SGML document is defined by: (1) a description of the structure of the document and (2) the text itself marked with tags which describe the structure.

Another promising direction to improve the effectiveness of search engines concerns the way in which it is possible to "filter" the great amount of information available across the Internet (Reisnick & Varian, 1997). Information filtering is a name used to describe a variety of processes involving the delivery of information to people who need it. The first filtering systems developed were based on document contents. However, it is known that more effective filtering can be done by involving humans in the filtering process. This idea is supported by the *collaborative filtering systems* or *recommender systems* (Reisnick & Varian, 1997). In these systems the people collaborate to help one another perform filtering by recording their reactions to documents they read. In a typical recommender system people provide evaluation judgements or annotations on documents as inputs, which the system then aggregates obtaining recommendations that directs to appropriate recipients. Later, these recommendations can be reused to assist another people in their search processes. Recommendations are a kind of plausible measures of the informative quality of documents. Usually, they are obtained according to a quantitative criterion, i.e., they require a critical number of distinct recommenders to be reached. On the other hand, in typical recommender systems is assumed that people express their evaluation judgements by means of numerical values. Sometimes, however a person could have a vague knowledge about judgement valuations, and cannot express his/her judgements with an exact numerical value. Then, a more realistic approach may be to use linguistic assessments to express the evaluation judgements instead of numerical values, i.e., to suppose that the variables which participate in the evaluation process are assessed by linguistic terms.

The *fuzzy linguistic approach* is a *soft computing tool* to manage linguistic information, which is based on the concept of *linguistic variable* (Zadeh, 1975a,b,c). It allows us to model in the problems qualitative values typical of human communication for representing qualitative concepts such as "importance" or "significance". Any fuzzy linguistic modeling needs adequate aggregation operators for *computing with words*. Two important operators for computing with words are the linguistic ordered weighted averaging (LOWA) operator (Herrera, Herrera-Viedma, & Verdegay, 1996) and the linguistic weighted averaging (LWA) operator (Herrera & Herrera-Viedma, 1997). They are designed to aggregate non-weighted and weighted linguistic information,

respectively. The fuzzy linguistic approach has been applied satisfactorily to different research areas. For example, in information retrieval to represent user queries and document relevance (Bordogna & Pasi, 1993; Delgado, Herrera, Herrera-Viedma, Martin-Bautista, & Vila, 2001; Herrera-Viedma, 2001a,b) or in Decision Making to represent user preferences and choice degrees of alternatives (Herrera et al., 1996). Particularly, in Fontana (2001) is presented an evaluation methodology of SGML-based documents using a fuzzy linguistic approach to represent the informative categories of documents. Then, given a linguistic informative category, an evaluation method based on fuzzy grammars is defined to obtain recommendations that state whether a document belongs to such a category or not (Fontana, 2001). This method presents the following limitations: (i) all components of DTD that participate in the evaluation process are equally important to obtain the recommendations, and this assumption is unrealistic, (ii) the recommenders' judgements are provided by numerical values, and (iii) the computational methods underlying associated to fuzzy grammars proposed (*inf-sup-based method, mean-sup-based method and mean–mean-based method*) present several drawbacks to take into account the majority of recommenders' evaluations. In this paper, we shall define a new evaluation method based on a fuzzy linguistic approach which overcomes the above limitations.

The main aim of the paper is to present a fuzzy soft computing method based on computing with words for evaluating the informative quality of documents in SGML format from judgements in order to generate recommendations. Given a kind of SGML-based document (e.g. "scientific article"), we establish an evaluation scheme composed by a subset of set of elements that define its DTD (e.g. "title, authors, abstract, introduction, body, conclusions, bibliography"). We assume that each component of that subset has a distinct informative role, i.e., each one affects the overall evaluation of a document in a different way. This peculiarity is added in the DTD by defining an attribute for each meaningful component that contains a relative linguistic importance degree. Then, given an area of interest (e.g. "Web publishing"), the recommendation for an SGML-based document is obtained by combining the linguistic evaluation judgements provided by different recommenders on the meaningful components of the document structure. To do so, we use an LWA–LOWA-based evaluation method. First, for each recommender we obtain his/her individual recommendation of document by aggregating his/her respective linguistic evaluation judgements weighted by relative linguistic importance degrees of evaluation scheme using the LWA operator. And then, we obtain a global recommendation of document by aggregating the individual recommendations using the LOWA operator. Both operators are guided by the concept of *fuzzy majority* (Herrera et al., 1996) represented by the *linguistic quantifiers* (Zadeh, 1983) used in the aggregations. In such a way, in the computation of the recommendations are taken into account the majority of recommenders' evaluation judgements, overcoming the drawback presented in Fontana (2001). The recommendations obtained are linguistic values that express qualitatively the informative quality of SGML-based documents with respect to an interest topic. With these recommendations the documents are arranged in linguistic informative categories. Finally, we should point out that an important advantage of this method is that with the addition of linguistic evaluation capabilities to SGML-based documents we are increasing the information filtering and evaluation possibilities in the Web.

The paper is set out as follows. The SGML is presented in Section 2. The fuzzy linguistic approach for computing with words is discussed in Section 3. The evaluation method of SGML-based documents is defined in Section 4. Finally, Section 5 includes our conclusions.

2. SGML-based documents

SGML is a metalanguage, that is, a means of formally describing a language. Specifically, SGML provides the rules for defining a markup language based on tags (Goldfarb, 1990). It has been developed to keep up the proliferation of proprietary formats in use for electronic document processing and representation. It is a descriptive system that gives a declarative and machine-independent description of the document structure using codes that simply offer names to categorize and identify the parts of a document. This means that SGML is a protocol devised to articulate structures of contents of documents instead of the appearance of documents.

2.1. Brief description of SGML

SGML introduces the notion of document type and, consequently, a DTD. The document type is formally defined by its constituent parts and structure. An SGML-based document structure is made up of a finite set of elements, each one associated to a (possible empty) finite set of attributes. The attributes are declarative specifications of the graphical rendering of the elements. Once a set of elements in SGML is defined for a given document, we have to give elements a syntactical structure. Such a structure is introduced in the form of a grammar through an DTD, i.e., by means of a finite set of declarative statements delimited by angle brackets of the form (Fontana, 2001):

<!ELEMENT *name min_rules content_model*>.

ELEMENT is a keyword of the SGML specifying that an element of document structure is being declared. *name* denotes the name of element. Each ELEMENT represents a tag denoted by *name*. *content_model* is a name of a string of elements that defines a syntactic structure for the element *name*. It is specified using a regular expression style syntax where "," stands for concatenation, "|" stands for logical or, "?" stands for zero or one occurrence, "$*$" stands for zero or more occurrences, and "+" stands for one or more occurrences of the preceding element. The *content_model* of an element can be composed of the combination of *content_model* of other elements, ASCII characters (PCDATA), binary data (NDATA), or EMPTY. And *min_rules* are the minimization rules of the element, given by an ordered pair of characters in the set $\{-, o\}$ that indicate if the starting and ending tags are compulsory (−) or optional (o). The possible attributes of an element are given in an attribute list (ATTLIST) identified by the element *name*, followed by the name of each attribute, its type, and if it is required or not (otherwise, the default value is given). Hence, an SGML document is defined by a DTD and the text itself marked with tags described in the DTD. Tags are denoted by angle brackets (<tagname>). Tags are used to identify the beginning and ending of pieces of the document. Ending tags are specified by adding a slash before the tag name (</tagname>). Tag attributes are specified at the beginning of the elements, inside the angle brackets and after the tagname using the syntax "attname = value".

Example 1. The following DTD involved by SGML represents the structure of a document that is a scientific article:

<!DOCTYPE article [
<!ELEMENT article - - (title, authors, abstract?, introduction, body, conclusions, bibliography)>
<!ELEMENT title - o (#PCDATA)>
<!ELEMENT authors - - (author+)>
<!ELEMENT (author | abstract | introduction) - o (#PCDATA)>
<!ELEMENT body - - (section+)>
<!ELEMENT section - o (titleS, #PCDATA)>
<!ELEMENT titleS - - (#PCDATA)>
<!ELEMENT conclusions - o (#PCDATA)>
<!ELEMENT bibliography - - (bibitem+)>
<!ELEMENT bibitem - o (#PCDATA)>]>

According to this DTD, the document "article" is composed by a title, at least an author, at most an abstract, an introduction, a body, a conclusions and a bibliography. The body is made up of at least one section and each section is composed by its respective title ("titleS") and characters. The bibliography is made up of at least one bibitem. The title, each author, abstract, introduction, each section title, conclusions and each bibitem is made up of characters.

Example 2. An example of a document instance of DTD defined in Example 1 may be the following:

<!DOCTYPE article SYSTEM "article.dtd">
<article>
<title>An Introduction to the Extensible Markup Language</title>
<authors>
<author>Martin Bryan</author> </authors>
<abstract>This article gives a very brief overview of the most commonly used components...
<introduction> XML was not designed to be a standardized way of coding text: in fact...
</introduction>
<body>
<section> <titleS>What is XML?</titleS> XML is subset of the Standard Generalized Markup Language (SGML) defined in ISO standard 8879:1986 that... </section>
<section><titleS>The components of XML</titleS> XML is based on the concept of documents composed of a series of ...
</body>
<conclusions> By storing data in the clearly defined format provided by XML you can ...
</conclusions>
<bibliography>
<bibitem>International Organization for Standardization. ISO 8879-1986 (E). Information Processing. Text and Office Systems. Standard Generalized Markup Language (SGML). Geneva: International Organization for Standardization, 1986.
</bibliography>
</article>

2.2. SGML, HTML or XML?

Some advantages of SGML are the following (Goldfarb, 1990; Larson, McDonough, O'Leary, & Kuntz, 1996): It facilitates accessibility by improving information discrimination, due to its capacity to structure contents. Additionally, in most cases it is easier and more effective to maintain an SGML database and translate the data into other formats according to current needs (including other SGML applications), than to maintain different copies in each of the formats needed. Its capability of including external and internal links to other documents maximizes browsing possibilities. And, finally, it is an appropriate tool for distinguishing between pertinent elements of information. Despite of this the benefits of SGML are not without cost and the manipulation of SGML object is impossible without specific software. Due to this, in the past few years, work on structured documents has centered on simplifying SGML; two of these efforts are HyperText Markup Language (HTML) and eXtensible Markup Language (XML) (Goldfarb & Prescod, 1998).

On the one hand, most documents on the Web are stored and transmitted in HTML. HTML is an instance of SGML, and although there is an HTML DTD, most HTML-based documents do not explicitly make reference to the DTD. HTML is a simple language well suited for hypertext, multimedia, and display of small and simple documents. However, HTML presents many limitations, e.g., it does not allow users to specify their own elements or attributes in order to semantically qualify their data.

On the other hand, XML is a simplified subset of SGML intended to make it more usable for distributing materials on the Web (Goldfarb & Prescod, 1998). XML is not a markup language, as HTML is, but a metalanguage that is capable of containing markup languages in the same way as SGML. The designers of XML simply took the best parts of SGML, guided by the experience with HTML, and produced something that is no less powerful than SGML, but vastly more regular and simpler to use. XML is a profile of SGML that eliminates many of the difficulties of implementing things (existence of a DTD), so for the most part it behaves just like SGML. Then, while SGML is mostly used for technical documentation and much less for other kinds of data, with XML it is exactly the opposite.

Consequently, in this paper we choose SGML as our reference structured format responding to the necessity of count with a global framework for the design of the "master" DTD easily adaptable to different vocabularies that may arise from SGML. Furthermore, in the context of information systems, keeping up a wide SGML data base and ad-hoc filtering to other representation formats (e.g. HTML and XML) has been a good development strategy that guarantees the reuse of research advances (Fausey & Shafer, 1997; Larson et al., 1996).

2.3. The evaluation problem of SGML-based documents

Assuming an SGML-based framework, an important interest area is the evaluation of informative quality of SGML documents to generate recommendations that allow to filter the information to help people in their search process. As we said at the beginning, the main evaluation and filtering systems developed involve human beings in their activity (Reisnick & Varian, 1997). People provide evaluation judgements or annotations on documents as inputs and the system aggregates them obtaining recommendations. In this context, a usual problem to solve is the

management of uncertainty and imprecision that appears in the evaluation judgements provided by the recommenders.

The fuzzy sets theory provides an adequate framework to manage the uncertainty and imprecision. In this sense, an evaluation methodology of SGML-based documents using a fuzzy approach was presented in Fontana (2001). This fuzzy evaluation methodology uses fuzzy grammars to evaluate the informative quality of SGML documents from evaluation judgements. As was pointed out in the introduction, it presents the following limitations: (i) all components of DTD that participate in the evaluation process are equally important to obtain the recommendations, and this assumption is unrealistic, (ii) the recommenders' judgements are provided by numerical values, and (iii) the computational methods underlying associated to fuzzy grammars proposed (*inf-sup-based method*, *mean-sup-based method and mean–mean-based method*) present several drawbacks to take into account the majority of recommenders' evaluations. Then, in order to overcome these limitations in Section 4 we propose a new evaluation methodology based on computing with words which incorporates linguistic evaluation capacities in the information evaluation and filtering systems of SGML documents. In the following section, we present a fuzzy linguistic approach for computing with words which is used to define that new evaluation methodology.

3. A fuzzy linguistic approach for computing with words

Many problems present fuzzy and unrigorous qualitative aspects (decision making, clinical diagnosis, information retrieval, etc). In such problems the information cannot be assessed precisely in a quantitative form but it may be done in a qualitative one, and thus, the use of a *linguistic approach* is necessary. The *fuzzy linguistic approach* is an approximate technique appropriate to deal with fuzzy and unrigorous qualitative aspects of problems (Herrera-Viedma, 2001a,b). It models linguistic information by means of linguistic terms supported by *linguistic variables* (Zadeh, 1975a,b,c). These are variables whose values are not numbers but words or sentences in a natural or artificial language. A linguistic variable is defined by means of a syntactic rule and a semantic rule. The fuzzy linguistic approach is less precise than the numerical one, but, however it presents the following advantages: (i) The linguistic description is easily understood by human beings even when the concepts are abstract or the context is changing, and furthermore, (ii) it diminishes the effects of noise since, as it is known the more refined assessment scale is, then more sensitive to noise (linguistic scales are less refined than numerical scales and consequently they are less sensitive to error apparition and propagation).

The *ordinal fuzzy linguistic approach* is a kind of fuzzy linguistic approach very useful and used for modeling the linguistic aspects in the problems. It facilitates the fuzzy linguistic modeling very much because it simplifies the definition of the semantic and syntactic rules (Herrera-Viedma, 2001a). In an ordinal fuzzy linguistic approach the syntactic rule is defined by considering a finite and totally ordered label set $S = \{s_i\}$, $i \in \{0, \ldots, \mathscr{T}\}$ in the usual sense, i.e., $s_i \geqslant s_j$ if $i \geqslant j$, and with odd cardinality. Typical values of cardinality used in the linguistic models, are odd values, 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 the rest of the terms being placed symmetrically around it. These classical cardinality values seems to fall in line with Miller's

observation about the fact that human beings can reasonably manage to bear in mind seven or so items (Miller, 1956). 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 $(s_i, s_{\mathcal{T}-i})$ is equally informative. Furthermore, to each label is associated a fuzzy number defined on the [0,1] interval, which is described by a linear trapezoidal membership function represented by the 4-tuple $(a_i, b_i, \alpha_i, \beta_i)$ (the first two parameters indicate the interval in which the membership value is 1; the third and fourth parameters indicate the left and right widths of the distribution).

Example 3. For example, we can use the following set of nine labels (Fig. 1) with its associated semantics in $U = [0, 1]$ to evaluate the linguistic variables used to provide the evaluations on documents:

$T = Total = (1, 1, 0, 0)$
$EH = Extremely\_High = (0.98, 0.99, 0.05, 0.01)$
$VH = Very\_High = (0.78, 0.92, 0.06, 0.05)$
$H = High = (0.63, 0.80, 0.05, 0.06)$
$M = Medium = (0.41, 0.59, 0.09, 0.07)$
$L = Low = (0.22, 0.36, 0.05, 0.06)$
$VL = Very\_Low = (0.1, 0.18, 0.06, 0.05)$
$EL = Extremely\_Low = (0.01, 0.02, 0.01, 0.05)$
$N = None = (0, 0, 0, 0)$

In any linguistic approach we need management operators of linguistic information. An advantage of the ordinal fuzzy linguistic approach is the simplicity and quickness of its computational model for computing with words. It is based on the *symbolic computation* (Delgado, Verdegay, & Vila, 1993; Herrera & Herrera-Viedma, 1997; Herrera et al., 1996). This technique acts by direct computation on labels by taking into account the order of such linguistic assessments in the ordered structure of linguistic terms. This symbolic tool seems natural when using the fuzzy linguistic approach, because the linguistic assessments are simply approximations which are given and handled when it is impossible or unnecessary to obtain more accurate values. Thus, in this case, the use of membership functions associated to the linguistic terms is unnecessary.

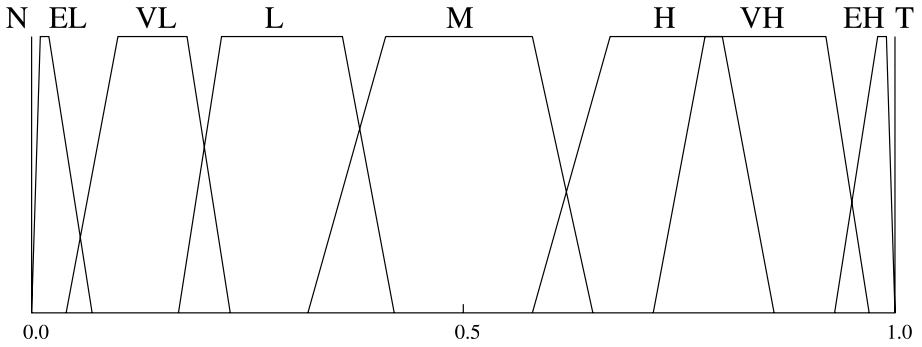

Fig. 1. A set of nine terms with its semantics.

Usually, the ordinal fuzzy linguistic model for computing with words is defined by establishing (i) a negation operator, (ii) comparison operators based on the ordered structure of linguistic terms, and (iii) adequate aggregation operators of ordinal fuzzy linguistic information.

In most ordinal fuzzy linguistic approaches the negation operator is defined from the semantics associated to the linguistic terms as $Neg(s_i) = s_j | j = \mathscr{T} - i$; and there are defined two comparison operators of linguistic terms:

1. Maximization operator: $\mathrm{MAX}(s_i, s_j) = s_i$ if $s_i \geqslant s_j$.
2. Minimization operator: $\mathrm{MIN}(s_i, s_j) = s_i$ if $s_i \leqslant s_j$.

In the following subsections, we present two aggregation operators to complete the ordinal linguistic computational model.

3.1. The LOWA operator

An important aggregation operator of ordinal linguistic values based on symbolic computation is the LOWA operator (Herrera et al., 1996). The LOWA operator is based on the *ordered weighted averaging* (*OWA*) *operator* defined in Yager (1988), and on the *convex combination of linguistic labels* defined in Delgado et al. (1993). It is used to aggregate non-weighted ordinal fuzzy linguistic information, i.e., linguistic information values with equal importance.

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

$$\begin{aligned}\phi(a_1, \ldots, a_m) &= W \cdot B^T = \mathscr{C}^m\{w_k, b_k, k = 1, \ldots, m\} \\ &= w_1 \odot b_1 \oplus (1 - w_1) \odot \mathscr{C}^{m-1}\{\beta_h, b_h, h = 2, \ldots, m\},\end{aligned}$$

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, $a_{\sigma(j)} \leqslant a_{\sigma(i)}\ \forall i \leqslant j$, with σ being a permutation over the set of labels A. $\mathscr{C}^m$ is the convex combination operator of m labels and if $m = 2$, then it is defined as

$$\mathscr{C}^2\{w_i, b_i, i = 1, 2\} = w_1 \odot s_j \oplus (1 - w_1) \odot s_i = s_k,$$

such that, $k = \min\{\mathscr{T}, i + round(w_1 \cdot (j - i))\} s_j, s_i \in S, (j \geqslant i)$, 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: $\mathscr{C}^m\{w_i, b_i, i = 1, \ldots, m\} = b_j$.

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

- $\phi(a_1, \ldots, a_m) = \mathrm{MAX}_i(a_i)$ if $W^* = [1, \ldots, 0]$,
- $\phi(a_1, \ldots, a_m) = \mathrm{MIN}_i(a_i)$ if $W_* = [0, \ldots, 1]$,
- $\phi(a_1, \ldots, a_m) = \mathrm{Ave}(a_i)$ if $W_A = [\frac{1}{m}, \ldots, \frac{1}{m}]$.

In order to classify OWA operators in regard to their location between *and* and *or*, Yager (1988) introduced a measure to characterize the type of aggregation being performed for a particular value of the weighting vector W. This measure, called the *orness measure* of the aggregation, is defined as

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

As suggested by Yager (1988) 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 (Yager, 1988) that $orness(W^*) = 1$, $orness(W_*) = 0$, and $orness(W_A) = 0.5$. 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 $orness(W)$. Therefore, an OWA operator with much of non-zero weights near the top will be an *orlike* operator ($orness(W) \geqslant 0.5$), and when much of the weights are non-zero near the bottom, the OWA operator will be *andlike* ($orness(W) < 0.5$).

An important question of the OWA operator is the determination of the weighting vector. A number of approaches have been suggested for obtaining the weights (Filev & Yager, 1998; Yager, 1993). A possible solution is that the weights represent the concept of fuzzy majority in the aggregation of LOWA operator using fuzzy linguistic quantifiers (Zadeh, 1983). Yager proposed an interesting way to compute the weights of the OWA operator, which, in the case of a non-decreasing proportional fuzzy linguistic quantifier, Q, is given by this expression (Yager, 1988, 1993): $w_i = Q(i/m) - Q((i-1)/m), i = 1, \ldots, m$, being the membership function of Q:

$$Q(r) = \begin{cases} 0 & \text{if } r < a \\ \frac{r-a}{b-a} & \text{if } a \leqslant r \leqslant b \\ 1 & \text{if } r > b \end{cases}$$

with $a, b, r \in [0,1]$. Some examples of non-decreasing proportional fuzzy linguistic quantifiers are: "most" $(0.3, 0.8)$, "at least half" $(0, 0.5)$ and "as many as possible" $(0.5, 1)$. When a fuzzy linguistic quantifier Q is used to compute the weights of LOWA operator, ϕ, it is symbolized by ϕ_Q.

3.2. The LWA operator

Another important aggregation operator of ordinal linguistic values is the LWA operator (Herrera & Herrera-Viedma, 1997). It is based on the LOWA operator and is defined to aggregate weighted ordinal fuzzy linguistic information, i.e., linguistic information values with not equal importance.

As it is known, the aggregation of weighted information involves two activities: (i) The transformation of the weighted information under the importance degrees by means of a transformation function h, and (ii) the aggregation of the transformed weighted information by means of an aggregation operator of non-weighted information f. The transformation function depends upon the type of aggregation of weighted information which is going to be performed. In Yager (1987), Yager 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 degree, which presents the non-increasing monotonic property in these importance degrees. For the MAX aggregation, he suggested a family of t-norms acting on weighted information and the importance degree, which presents the non-decreasing monotonic property in these importance degrees.

Following the above ideas, we define the LWA operator in (Herrera & Herrera-Viedma, 1997). Here, we redefine it to simplify its expression using the orness measure and as f the LOWA operator ϕ.

Definition 2. The aggregation of a set of weighted linguistic opinions, $\{(c_1, a_1), \ldots, (c_m, a_m,)\}$, $c_i, a_i \in S$, according to the LWA operator Φ is defined as

$$\Phi[(c_1, a_1), \ldots, (c_m, a_m)] = \phi(h(c_1, a_1), \ldots, h(c_m, a_m)),$$

where a_i represents the weighted opinion, c_i the importance degree of a_i, and h is the transformation function defined depending on the weighting vector W assumed for the LOWA operator ϕ, such that,

$$h = LC_v^{\rightarrow} \text{ if } orness(W) \geqslant 0.5, \quad \text{and} \quad h = LI_v^{\rightarrow} \text{ if } orness(W) < 0.5,$$

where $LC_v^{\rightarrow}$ are the following group of linguistic t-norms, called the linguistic conjunction functions, which are monotonically non-decreasing in the weights:

1. The classical MIN operator: $LC_1^{\rightarrow}(c, a) = \text{MIN}(c, a)$.
2. The nilpotent MIN operator: $LC_2^{\rightarrow}(c, a) = \begin{cases} \text{MIN}(c, a) & \text{if } c > Neg(a) \\ s_0 & \text{otherwise.} \end{cases}$
3. The weakest conjunction: $LC_3^{\rightarrow}(c, a) = \begin{cases} \text{MIN}(c, a) & \text{if } \text{MAX}(c, a) = s_{\mathscr{T}} \\ s_0 & \text{otherwise.} \end{cases}$

and where $LI_k^{\rightarrow}$ are the following group of linguistic implication functions, called the linguistic implication functions, which are monotonically non-increasing in the weights:

1. Kleene–Dienes's implication function: $LI_1^{\rightarrow}(c, a) = \text{MAX}(Neg(c), a)$.
2. Gödel's implication function: $LI_2^{\rightarrow}(c, a) = \begin{cases} s_{\mathscr{T}} & \text{if } c \leqslant a \\ a & \text{otherwise.} \end{cases}$
3. Fodor's implication function: $LI_3^{\rightarrow}(c, a) = \begin{cases} s_{\mathscr{T}} & \text{if } c \leqslant a \\ \text{MAX}(Neg(c), a) & \text{otherwise.} \end{cases}$

Remark 1. We should point out that the LOWA and LWA operators are the basis of new linguistic evaluation model of SGML-based documents that we present in this paper. We have chosen these operators due to the following reasons: (i) both operators are complementary (the LWA operator is defined from the LOWA operator) and this simplifies the design of evaluation method, (ii) both operators act by symbolic computation and therefore linguistic approximation processes are unnecessary and this simplifies the processes of computing with words, and finally (iii) the concept of fuzzy majority represented by linguistic quantifiers acts in their processes of

computation and, in such a way, the recommendations on documents are obtained according to the majority of evaluations provided by the panel of recommenders, overcoming the limitations of evaluation model presented in Fontana (2001).

4. Evaluating SGML-based documents for generating recommendations

In this section we present a method for evaluating the informative quality of documents described in SGML format with the aim of assigning them linguistic recommendation values. The linguistic recommendations are obtained from the linguistic evaluation judgements provided by a non-determined number of recommenders on the more important elements of DTD considered. They are achieved applying the *LWA–LOWA-based evaluation method* developed for computing with words.

4.1. Evaluation procedure of the SGML-based documents

Suppose that we want to generate a recommendation database for qualifying the information of a set of SGML-based documents $\{d_1, \ldots, d_l\}$ with the same DTD. These documents can be evaluated from a set of different areas of interest, $\{\mathscr{A}_1, \ldots, \mathscr{A}_q\}$. Consider an evaluation scheme composed by a finite number of elements of DTD, $\{p_1, \ldots, p_n\}$, which will be evaluated in each document d_k by a panel of recommenders or referees $\{e_1, \ldots, e_m\}$. We assume that each component of that evaluation scheme presents a distinct informative role. This is modeled by assigning to each p_j a relative linguistic importance degree $I(p_j)$ supported by the linguistic variable "*Importance*" defined as in Section 3, i.e., $I(p_j) \in S = \{s_0, s_1, \ldots, s_{\mathscr{T}}\}$. Each importance degree $I(p_j)$ is a measure of the relative importance of element p_j with respect to others existing in the evaluation scheme. We propose to include these relative linguistic importance degrees in the DTD. This can be done easily by defining in the DTD an attribute of importance "rank" for each component of evaluation scheme using the SGML syntax.

Example 4. For example in the DTD given in Example 1, supposing that the evaluation scheme is composed by the elements (title, authors, abstract, introduction, body, conclusions, bibliography), the importance degrees can be considered by adding the following declarative statement:

<!ATTLIST (title | authors | abstract | introduction | body | conclusions | bibliography) rank ($s_0|s_1|s_2|\ldots|s_{\mathscr{T}-1}|s_{\mathscr{T}}$)#*REQUIRED*>

Then, the instance of document given in Example 2 would be as follows:

<!DOCTYPE article SYSTEM "article.dtd">
<article>
<title rank = "I(title)">An Introduction to the Extensible Markup Language</title>
<authors rank = "I(authors)">
<author>Martin Bryan</author> </authors>
<abstract rank = "I(abstract)">This article gives a very brief overview of the most commonly used components...

<introduction rank = "I(introduction)"> XML was not designed to be a standardized way of coding text: in fact...</introduction>
<body rank = "I(body)">
<section> <titleS>What is XML?</titleS> XML is subset of the Standard Generalized Markup Language (SGML) defined in ISO standard 8879:1986 that... </section>
<section><titleS>The components of XML</titleS> XML is based on the concept of documents composed of a series of ...
</body>
<conclusions rank = "I(conclusions)"> By storing data in the defined format ... </conclusions>
<bibliography rank = "I(bibliography)">
<bibitem>International Organization for Standardization. ISO 8879-1986 (E). Information Processing. Text and Office Systems. Standard Generalized Markup Language (SGML). Geneva: International Organization for Standardization, 1986.
</bibliography>
</article>

Let e_{kt}^{ij} be linguistic evaluation judgement provided by the recommender e_k measuring the informative quality or significance of element p_j of document d_i with respect to the area of interest $\mathscr{A}_t$. Consider that e_{kt}^{ij} is supported by the linguistic variable "Significance", which uses the same label set associated to "Importance", but with a different interpretation, i.e., $e_{kt}^{ij} \in S$. Then, the evaluation procedure of a SGML-based document d_i obtains a recommendation $r_t^i \in S$ (it is also supported by the linguistic variable "Significance") using the LWA–LOWA-based evaluation method in the following steps:

1. Capture the topic of interest ($\mathscr{A}_t$), the linguistic importance degrees of evaluation scheme fixed in the DTD $\{I(p_1), \ldots, I(p_n)\}$, and all the evaluation judgements provided by the panel of recommenders $\{e_{kt}^{ij}, j = 1, \ldots, n\}, k = 1, \ldots, m$.
2. Calculate for each e_k his/her individual recommendation r_{kt}^i by means of the LWA operator as

$$r_{kt}^i = \Phi[(I(p_1), e_{kt}^{i1}), \ldots, (I(p_n), e_{kt}^{in})] = \phi_{Q_2}(h(I(p_1), e_{kt}^{i1}), \ldots, h(I(p_n), e_{kt}^{in})).$$

   Therefore, r_{kt}^i is a significance measure that represents the informative quality of d_i with respect to topic $\mathscr{A}_t$ according to the Q_2 evaluation judgements provided by e_k.
3. Calculate the global recommendation r_t^i by means of an LOWA operator guided by the fuzzy majority concept represented by a linguistic quantifier Q_1 as

$$r_t^i = \phi_{Q_1}(r_{1t}^i, \ldots, r_{mt}^i).$$

   In this case, r_t^i is a significance measure that represents the informative quality of d_i with respect to topic $\mathscr{A}_t$ according to the Q_2 evaluation judgements provided by the Q_1 recommenders. r_t^i represents the linguistic informative category of d_i with respect to the topic $\mathscr{A}_t$.
4. Store the recommendation r_t^i in a recipient in order to assist users in their later search processes.

In the evaluation procedure the linguistic quantifiers Q_1 and Q_2 represent the concept of fuzzy majority in the computing process with words. In such a way, the recommendations on documents

are obtained by taking into account the majority of evaluations provided by the majority of recommenders.

Remark 2. We should point out that the recommendation value obtained in the evaluation procedure can be stored in a file "significance.dat", which can be considered in the DTD using "entity references". For example, in the DTD presented in Example 1 we can add the following entity declaration that identifies the file that contains the recommendation:

<!ENTITY %Si SYSTEM "significance.dat">

And then, by defining an attribute to store the recommendation in each document instance with the following declarative statement:

<!ATTLIST article significance %Si; #REQUIRED.>

Remark 3. We can extend easily this evaluation method for evaluating XML-based documents since XML is also a metalanguage although more restricted. Moreover, this is convenient because XML is more used to serving documents over the Web than SGML, and in such a way, we can include the evaluation model across the Web. To do so, we have to claim XML-based documents the use of DTD. Indeed, the use of a DTD assure the convergence between SGML/XML and future extensions to HTML, because the extension of method to HTML is also easy. To apply the method to XML-based documents we have to define the DTD following the rules for XML and furthermore, to include in the DTD the requirement to use DTD with the following declarative statement:

<?xml version = "1.0" encoding = "UTF-16" standalone = "yes" rmd = "internal"?>

In this manner every XML document instance bring into the matter the necessary pair attribute-value in the corresponding element.

4.2. Example of application

Fixed a topic of interest, e.g., $\mathscr{A}_t$ = "webpublishing", we evaluate the informative quality of SGML document d_i presented in Example 2 according to the evaluation scheme with seven components assumed in Example 4, i.e., ("title, authors, abstract, introduction, body, conclusions, bibliography").

Consider the set of nine labels given in Example 3, $S = \{N, EL, L, M, H, VH, EH, T\}$, to express the linguistic information. Suppose that the evaluation scheme presents the following relative linguistic importance degrees: $\{\mathrm{I}(\text{title}) = H, \mathrm{I}(\text{authors}) = L, \mathrm{I}(\text{abstract}) = VH, \mathrm{I}(\text{introduction}) = VH, \mathrm{I}(\text{body}) = T, \mathrm{I}(\text{conclusions}) = H, \mathrm{I}(\text{bibliography}) = L\}$. Then, assume that a group of four recommenders evaluate the components of evaluation scheme providing the following linguistic evaluation judgements: $e_1 : (H, L, M, H, M, L, VL)$, $e_2 : (M, VL, H, H, M, L, VL)$, $e_3 : (L, L, L, M, M, L, VL)$, $e_4 : (H, N, M, VH, M, L, L)$.

Suppose the linguistic quantifier "*at least half*" with the pair (0,0.5) to represent the concept of fuzzy majority in the LWA–LOWA-based evaluation method. With this quantifier we have the following weighting vector $W_2 = (0.28, 0.28, 0.28, 0.16, 0, 0, 0)$ for calculating the linguistic individual recommendations with the LWA operator. As $orness(W_2) = 0.78 \geqslant 0.5$, then LOWA operator used in the aggregation of LWA operator is an orlike, and thus, we can choose as transformation function h any linguistic conjunction function LC_v, for example, $LC_1 = \text{MIN}$. Then, aggregating the linguistic evaluation judgements by means of the LWA operator we obtain the following linguistic individual recommendations: $r^i_{1t} = M$, $r^i_{2t} = M$, $r^i_{3t} = L$, $r^i_{4t} = H$. For example the value r^i_{4t} is obtained as

$$\begin{aligned} r^i_{4t} &= \Phi[(H,H),(L,N),(VH,M),(VH,VH),(T,M),(H,L),(L,L)] \\ &= \phi_{Q_2}(\text{MIN}(H,H),\text{MIN}(L,N), \\ &\text{MIN}(VH,M),\text{MIN}(VH,VH),\text{MIN}(T,M),\text{MIN}(H,L),\text{MIN}(L,L)) \\ &= \mathscr{C}^7\{VH,0.28,H,0.28,M,0.28,M,0.16,L,0,L,0,N,0\} \\ &= 0.28 \odot VH \oplus 0.72 \odot \mathscr{C}^6\{H,0.39,M,0.39,M,0.22,L,0,L,0,N,0\}, \end{aligned}$$

then if we develop the recursive definition of the LOWA

$$\begin{aligned} &\mathscr{C}^6\{H,0.39,M,0.39,M,0.22,L,0,L,0,N,0\} \\ &\quad = 0.39 \odot H \oplus 0.61 \odot \mathscr{C}^5\{M,0.64,M,0.36,L,0,L,0,N,0\}, \\ &\mathscr{C}^5\{M,0.64,M,0.36,L,0,L,0,N,0\} = 0.64 \odot M \oplus 0.36 \odot \mathscr{C}^4\{M,1,L,0,L,0,N,0\}, \\ &\mathscr{C}^4\{M,1,L,0,L,0,N,0\} = M \end{aligned}$$

and then evaluating by a bottom up process

$$\mathscr{C}^5\{M,0.64,\ldots,N,0\} = 0.64 \odot M \oplus 0.36 \odot M = s_4 = M,$$

given that $4 = \min\{\mathscr{T} = 8, 4 + round(0.64 \cdot (4-4))\}$. Then

$$\mathscr{C}^6\{H,0.39,\ldots,N,0\} = 0.39 \odot H \oplus 0.61 \odot M = M$$

and consequently

$$\mathscr{C}^7\{VH,0.28,\ldots,N,0\} = 0.28 \odot VH \oplus 0.72 \odot M = H.$$

Finally, using the LOWA operator with the weighting vector $W_1 = (0.5, 0.5, 0, 0)$ we obtain the following linguistic global recommendation: $r^i_t = H$. Therefore, the document d_i presents high value of significance and it can be perfectly recommended when users need documents related to the topic "Web publishing."

5. Conclusions

In this paper, we have presented a fuzzy linguistic evaluation method to characterize the information contained in SGML-based documents. The method generates linguistic recommendations for structured documents by taking into account the fuzzy majority of linguistic evaluation judgements provided by different recommenders to evaluate the informative quality of

the more meaningful component of DTD. The use of fuzzy linguistic modeling facilitates the activity of the filtering systems due to that the user–system interaction is more user-friendly.

The method proposed can be applicable to any DTD and also can be extended by considering other aspects in the generation of the recommendations, e.g., *aesthetic qualities*, *all multimedia information* (*including even text embedded in images*), *network factors such as loading time*, *clarity*, *originality*, *organization*, *etc*.

SGML is an standard non-proprietary format to represent the content of documents which increases considerably the retrieval and processing possibilities of information stored in the Web. With the addition of linguistic evaluation capabilities to SGML-based documents we increases also the information filtering and evaluation possibilities of information in the Web. Although SGML is not very used in the Web publishing, however this approach can be used with HTML documents in the Web due to it's possible to work with XML code in an HTML page using the *data-binding technique* and vice versa, using the *namespace specification* with the HTML code in an XML document. Similarly, this method can be applied also to XML-based documents, since XML is a much-restricted form of SGML and any fully conformed SGML system will be able to read XML documents.

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.

Delgado, M., Herrera, F., Herrera-Viedma, E., Martin-Bautista, M. J., & Vila, M. A. (2001). Combining linguistic information in a distributed intelligent agent model for information gathering on the internet. In P. P. Wang (Ed.), *Computing with words* (pp. 275–294). NY: John Wiley & Sons.

Delgado, M., Verdegay, J. L., & Vila, M. A. (1993). On aggregation operations of linguistic labels. *International Journal of Intelligent Systems, 8*, 351–370.

Fausey, J., & Shafer, K. (1997). All my data is in sgml. Now what? *Journal of the American Society for Information Science, 48*(7), 638–643.

Filev, D., & Yager, R. R. (1998). On the issue of obtaining OWA operator weights. *Fuzzy Sets and Systems, 94*, 157–169.

Fontana, F. A. (2001). Evaluation of SGML-based information through fuzzy techniques. *Information Processing & Management, 37*, 75–90.

Goldfarb, C. (1990). *The SGML handbook*. Oxford: Oxford University Press.

Goldfarb, C., & Prescod, P. (1998). *The XML handbook*. Oxford: Prentice Hall.

Herrera, F., & Herrera-Viedma, E. (1997). Aggregation operators for linguistic weighted information. *IEEE Transactions on Systems, Man, and Cybernetics Part A, 27*, 646–656.

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. (2001a). 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.

Herrera-Viedma, E. (2001b). An information retrieval system with ordinal linguistic weighted queries based on two weighting elements. *International Journal Uncertainty, Fuzziness and Knowledge-Based Systems, 9*, 77–88.

Kobayashi, M., & Takeda, K. (2000). Information retrieval on the web. *ACM Computing Surveys, 32*(2), 144–173.

Larson, R. R., McDonough, J., O'Leary, P., & Kuntz, L. (1996). Cheshire II: Designing a next-generation online catalog. *Journal of the American Society for Information Science, 47*(7), 555–567.

Lawrence, S., & Giles, C. (1998). Searching the web: General and scientific information access. *IEEE Communications Magazine, 37*(1), 116–122.

Miller, G. A. (1956). The magical number seven or minus two: Some limits on our capacity of processing information. *Psychological Review, 63*, 81–97.

Reisnick, P., & Varian, H. R. (1997). Recommender systems. *Special issue of Communications of the ACM, 40*(3).

Wium, H. L., & Saarela, J. (1999). Multipurpose web publishing using HTML, XML, and CSS. *Communications of the ACM, 42*(10), 95–101.

Yager, R. R. (1987). A note on weighted queries in information retrieval systems. *Journal of the American Society for Information Science, 38*, 23–24.

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

Yager, R. R. (1993). Families of OWA operators. *Fuzzy Sets and Systems, 59*, 125–148.

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

Zadeh, L. A. (1975b). The concept of a linguistic variable and its applications to approximate reasoning. *Part II. Information Sciences, 8*, 301–357.

Zadeh, L. A. (1975c). The concept of a linguistic variable and its applications to approximate reasoning. *Part III. Information Sciences, 9*, 43–80.

Zadeh, L. A. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computers and Mathematics with Applications, 9*, 149–184.