Eduardo Peis[1], Enrique Herrera-Viedma[2], Juan Carlos Herrera[2]
[1] Dept. Library & Information Science. University of Granada. Granada
[2] Dept. of Computer Science and A.I. University of Granada. 18071-Granada

On Evaluation of XML Documents Using Fuzzy Linguistic Techniques

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 XML documents based on computing with words is presented. Given an XML document type (e.g. scientific article), we consider that its elements are not equally informative. This is indicated by the use of a DTD and defining linguistic importance attributes to the more meaningful elements of the DTD designed. Then, the evaluation method generates linguistic recommendations from linguistic evaluation judgements provided by different recommenders on meaningful elements of DTD.

1. Introduction

Finding relevant, high quality information on the World Wide Web (WWW) is a difficult task. The 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 and Takeda, 2000), (Lawrence and Giles, 1998).

The development of standard formats for the representation of documents in Web improves substantially the quality of information retrieved by search engines. The logic structure of the documents on the web can be expressed with metalanguages like XML (Goldfarb and Prescod, 1998). The *eXtensible Markup Language* (XML) is a simplified subset of the *Standard Generalized Markup Language* (SGML) intended to make it more usable for distributing materials on the Web. XML is not a markup language, as *HyperText Markup Language* (HTML, which is another well-known subset of SGML) 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 and produce something that is no less powerful than SGML, but vastly more regular and simpler to use.

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. 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 and Varian, 1997). Usually, recommendations 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 main aim of the paper is to present a fuzzy soft computing method based on *computing with words* (Herrera et al., 1996) for evaluating the informative quality of documents in XML format in order to generate recommendations. The recommendations are obtained from the evaluation judgements provided by a panel of selected recommenders using a computing method based on the LWA (Herrera and Herrera-Viedma, 1997) and LOWA (Herrera et al., 1996) operators. The recommendations are linguistic values that express qualitatively the informative quality of XML based documents with respect to an interest topic. With these recommendations the documents are arranged in linguistic informative categories and, in such a way, later they can be reused easily to assist another people in their search processes.

To do so, the paper is structured as follows. The XML is presented in Section 2. Section 3 is devoted to introduce the tools of computing with words. Then, the evaluation method of XML documents is defined in Section 4. Finally, several conclusions are pointed out in Section 5.

2. XML Based Documents

XML is a subset of SGML, but while SGML is mostly used for technical documentation and much less for other kinds of data, with XML it is exactly the opposite, being it more usable for distributing materials on the Web (Goldfarb and Prescod, 1998).

Therefore, as SGML, XML provides the rules for defining a markup language based on tags. 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 XML is a protocol devised to articulate structures of contents of documents instead of the appearance of documents.

XML allows for the creation of custom tokens and custom document structures. Each XML document and each element of an XML document is an object with its own properties. The main difference between SGML and XML is that many XML based documents don't need an DTD. In our case, for representing the different evaluation variables we have worked with XML valid document (i.e. XML well formed document with correspondence, instead, to a DTD). A DTD serves as a template that helps to explain syntax and content of a document that is based on a specific DTD. Once a set of tokens is defined for a given document, we have to give tokens a syntactical structure. Such a structure is introduced in the form of a grammar in the DTD by means of a finite set of declarative statements delimited by angle brackets of the form:

<!ELEMENT *name content_model* >.

ELEMENT is a keyword specifying that an element or token 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*. 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 XML document can be 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 XML represents the structure of a document that is a scientific article:

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

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

```
<?xml version="1.0" standalone="no" ?>
 <!DOCTYPE article SYSTEM "article.dtd">
<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....
</abstract>
<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 ... </section></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>

3. Tools of Computing with Words

A *fuzzy linguistic approach* is an approximate technique appropriate to deal with qualitative aspects of problems (Zadeh, 1975). An *ordinal fuzzy linguistic approach* (Herrera et al., 1996) is a kind of fuzzy linguistic approach very useful and used for modelling the processes of computing with words and the linguistic aspects in the problems. It is defined by considering a finite and totally ordered label set, $S = \{s_i, i \in H = 0, \ldots, T\}, s_i \geq s_j, if \; i \leq j,$ in the usual sense, and with odd cardinality (7 or 9 labels). The mid term representing an assessment of "approximately 0.5" and the rest of the terms being placed symmetrically around it. The semantic of the linguistic term set is established from the ordered structure of the term set by considering that each linguistic term for the pair *(s_i, s_{T-i})* is equally informative.

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* (Herrera and 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.

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.

$$1.- There \; is \; the \; negation \; operator: Neg(s_i) = s_j, \; with \; j = T - i.$$

$$2.- Maximization \; operator: MAX(s_i, s_j) = s_i \; if \; s_i \geq s_j.$$

$$3.- Minimization \; operator: MIN(s_i, s_j) = s_i \; if \; s_j \leq s_j.$$

In the following subsections, to complete the ordinal linguistic computational model we present two aggregation operators that we shall use to define the evaluation method of XML documents.

3.2. The LOWA Operator

The *Linguistic Ordered Weighted Averaging* (LOWA) is an aggregation operator of ordinal linguistic values based on symbolic computation (Herrera et al., 1996). It is used to aggregate non-weighted ordinal linguistic information, i.e., linguistic information values with equal importance.

Definition 1. *Let* $A = \{ a_1, . . . , a_m \}$ *be a set of labels to be aggregated, then the LOWA operator,* Φ, *is defined as* $\Phi (a_1, . . . , a_m) = W \cdot B^T = C^m\{ w_k, b_k, k = 1, . . . , m \} = w_1 \Theta b_1 \oplus (1 - w_1) \Theta C^{m - 1} \{\beta_h, b_h, h = 2, . . . , m \}$, *where* $W = [w_1, . . . , w_m]$, *is a weighting vector, such that,* $w_i \in [0, 1]$ *and* $\sum_i w_i =1$. $\beta_h = w_h/(\sum_2^m \ w_{k)}$, $h =2, . . . , m$, *and* $B = \{ b_1, . . . , b_m \}$ *is a vector associated to A, such that,* $B = \sigma(A) = \{ a_{\sigma(1)} , . . . , a_{\sigma(m)}\}$, *where,* $a_{\sigma (j)} \leq a_{\sigma (i)} \ \forall \ i \leq j$, *with* σ *being a permutation over the set of labels A.* C^m *is the convex combination operator of m labels and if m=2, then it is defined as* $C^2\{ w_i, b_i, i = 1, 2 \} = w_1 \ \Theta \ s_j \oplus (1 - w_1) \ \Theta \ s_i = sk$, *such that* $k = \min \{ T, i + \text{round } (w_1 \cdot (j - i)) \}$ $s_j , s_i \in S$, $(j \geq i)$, *being "round" the usual round operation, and* $b_1 = s_j$, $b_2 = s_i$. *If* $w_j = 1$ *and* $w_i = 0$ *with* $i \neq j \forall i$, *then* $C^m\{ w_i, b_i, i = 1, . . . , m \} = b_j$.

The LOWA operator is an *"or-and"* operator (Herrera et al., 1996). This property allows that the LOWA operator carries out a soft computing in the modelling of MAX and MIN linguistic operators. In order to classify OWA operators in regard to their localisation between *and* and *or*, Yager (Yager, 1988) introduced a measure of *orness,* associated with any vector W as follows

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

An important question of the LOWA operator is the determination of W. A possible solution consists of representing the concept of *fuzzy majority* by means of the weights of W, using a non-decreasing proportional *fuzzy linguistic quantifier* (Zadeh, 1983), Q, in its computation (Yager, 1988): $w_i = Q(i/m) - Q((i\text{-}1)/m)$, , $i = 1, . . . , m$, being the membership function of Q,

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

With a, b, r $\in$ [0,1]. Sompe examples of non-decreasing proportional fuzzy linguistic quantifier are “most” (0.3, 0.8), “at least half” (0, 0.5) and “as many as possible” (0.5, 1).

3.3. The LWA Operator

The *Linguistic Weighted Averaging (LWA)* operator (Herrera and Herrera-Viedma, 1997) is another important aggregation operator which is based on the LOWA operator. It is defined to aggregate weighted ordinal linguistic information, i.e., linguistic information values with not equal importance.

Definition 2. *The aggregation of a set of weighted linguistic opinions,* $\{(c_{1,} a_1), . . . ,(c_{m,} a_m) \}$ $c_i , a_i \in S$, *according to the LWA operator* Π *is defined as* $\Pi[(c_{1,} a_1), . . . ,(c_{m,} a_m)] = \Phi(h(c_{1,} a_1), . . . ,h(c_{m,} a_m))$, *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,* $h= MIN(c_i , a_i)$ *if* $orness(W) \geq 0.5$, *and* $h= MAX(Neg(c_i), a_i)$ *if* $orness(W) < 0.5$.

4. Evaluating XML Based Documents for Generating Recommendations

Suppose that we want to generate a recommendation database for qualifying the information of a set of valid XML based documents, $\{d_1, . . ., d_l \}$, with the same DTD. These documents can be evaluated from a set of different areas of interest, $\{A_1 ,..., A_q \}$. Consider an evaluation scheme composed by a finite number of elements of DTD, $\{p_1 ,..., p_n \}$, which will be evaluated in each document d_k by a panel of recommenders or referees $\{e_1 ,..., e_m \}$. 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 2, i.e., $p_j \in S=\{s_1 ,..., s_T \}$. 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 XML syntax.

Let e^{ij}_{kt} be a 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 A_t. Consider that e^{ij}_{kt} is supported by the linguistic variable "Significance", which uses the same label set associated to "Importance", but with a different interpretation, i.e., $e^{ij}_{kt} \in S$. Then, the evaluation procedure of a XML based document d_i obtains a recommendation ., $r^{i}_{t} \in S$. (it is also supported by the linguistic variable "Significance") using evaluation method based on the LWA and LOWA operators which is composed by the following steps:

1. Capture the topic of interest A_t, the linguistic importance degrees of evaluation scheme fixed in the DTD $\{I(p_1),..., I(p_n)\}$, and all the evaluation judgements provided by the panel of recommenders $\{ e^{ij}_{kt}\}, j=1,...,n, k=1,...,m$.
2. Calculate for each e_k his/her individual recommendation r^{i}_{kt} by means of the LWA operator as
   $$r^{i}_{kt} =\Pi[(I(p_1), e^{i1}_{kt}) ,...,(I(p_n), e^{in}_{kt})]=)]=\Phi(h(I(p_1), e^{i1}_{kt}),..., h(I(p_n), e^{in}_{kt})).$$
   Therefore, r^{i}_{kt} is a significance measure that representsthe informative quality of d_i with respect to topic A_t according to the Q evaluation judgements provided by e_k , being Q the linguistic quantifier used to compute the weighting vector of Φ.

3. Calculate the global recommendation r^{i}_{t} by means of Φ guided by the fuzzy majority concept represented by the linguistic quantifier Q as $r^{i}_{t} =\Phi(r^{i}_{1t}, ..., r^{i}_{mt})$. Then , r^{i}_{t} is a significance measure that represents the informative quality of d_i with respect to A_t according to the Q evaluation judgements provided by the Q recommenders. r^{i}_{t} represents the linguistic informative category of d_i with respect to A_t .
4. Store the recommendation r^{i}_{t} in a recipient in order to assist users in their later search processes.

5. Conclusions

In this paper, we have presented a fuzzy linguistic evaluation method to characterize the information contained in XML 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.

References

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

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

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.

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

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

Reisnick, P. & Varian, H.R. (1997), Special issue on recommender systems. *Communications of the ACM*, 40(3): 56–89.

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

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

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