PRE-RANKING DOCUMENTS VALORIZATION IN THE INFORMATION RETRIEVAL PROCESS

Chkiwa Mounira[1], Jedidi Anis[1] and Faiez Gargouri[1]

[1]Multimedia, InfoRmation systems and Advanced Computing Laboratory
Sfax University, Tunisia
m.chkiwa@gmail.com, jedidianis@gmail.com, faiez.gargouri@isimsf.rnu.tn

ABSTRACT

In this short paper we present three methods to valorise score relevance of some documents basing on their characteristics in order to enhance their ranking. Our framework is an information retrieval system dedicated to children. The valorisation methods aim to increase the relevance score of some documents by an additional value which is proportional to the number of multimedia objects included, the number of objects linked to the user particulars and the included topics. All of the three valorization methods use fuzzy rules to identify the valorization value.

KEYWORDS

Information Retrieval, Pre-ranking valorization, Fuzzy Logic, Fuzzy Rules

1. INTRODUCTION

In the information retrieval process, younger users have particularities concerning what they really need in results. In this paper we study those particularities in order to have maximum coverage of relevant documents. To do it, we present the Pre-ranking Documents Valorization which aims to increase some documents relevance score by an additional value in order to enhance their ranking. In order to make the additional value be proportional to the document characteristics we use fuzzy logic principles. The pre-ranking document valorization takes place after running the querying process which finds the relevant documents. Our framework materializes collaboration between two axes: the Semantic Web [1 and 2] and the Fuzzy Logic [3, 4 and 5] .We use semantic web technologies as RDF [6 and 7] to annotate semantically our collection of web documents and SPARQL [8] for querying the annotation. Also, we use Fuzzy rules to find the exact value added to a score relevance in order to enhance the ranking of the correspondent document. The rest of the paper is organized as follows: in section 2 we introduce some preliminaries about our framework which is an information retrieval system dedicated for children. Section 3 we present the pre-ranking document valorization by explaining its three types. Finally section 4 concludes the paper.

David C. Wyld et al. (Eds) : CST, ITCS, JSE, SIP, ARIA, DMS - 2014
pp. 353–359, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4132

2. FRAMEWORK

To ensure a high quality of use, available information retrieval systems dedicated for children have common highlighted facts:

- Security: Since the web covers a large amount of uncontrollable data, security represents the first factor taken into account to create a safe information retrieval system dedicated for children.
- Design: it represents the main visual factor to call users attention.
- Profile: it represents the common way used to personalize a search process taking into account the user's personal interests.
- Querying: this factor makes the difference between information search engines even if it follows outwardly the same demarche: the annotation, the query/document matching, the ranking and the result display.

In addition of considering the cited facts, our particularity resides in the "pre-ranking document valorisation" which is localized in the figure below.

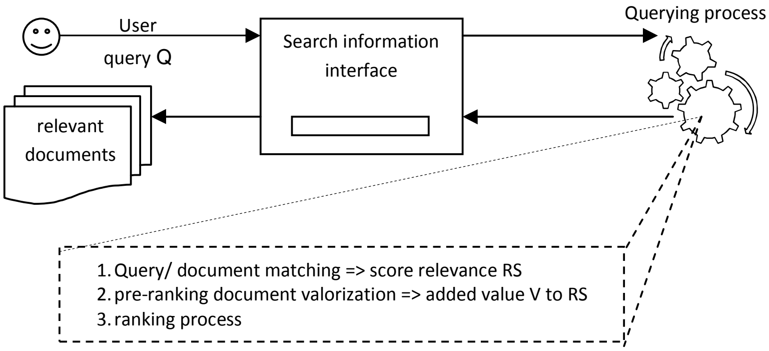
User
query Q
Search information interface
Querying process
relevant documents
1. Query/ document matching => score relevance RS
2. pre-ranking document valorization => added value V to RS
3. ranking process

Figure 1. Our querying particularity

In order to explore young web users searching interests, we do a survey covering 723 children between 8 and 15 years. The results show that more than 75% of them prefer results rich by multimedia objects (especially pictures and video sequences). Likewise, more than 70% of surveyed children want to see in returned results information dealing with their own particulars (own country, avenue, friends …) especially with the success of social networks use. In the next section we present two flexible methods that give the priority to relevant results rich by multimedia objects and/or dealing with the current user particulars. In the other side, we consider the fact that a returned document may be exhaustive and deal with different topics; this fact could misplace a user attention especially if it is a child. We propose therefore a method that avoid "multi-topic" documents and make priority to documents focusing only on query main topic. The three pre-ranking valorization methods are submitted to fuzzy rules which mainly use variables representing the user ages; Figure 2 shows the membership function representing web young user's age as input set which ranges from 3 to 14

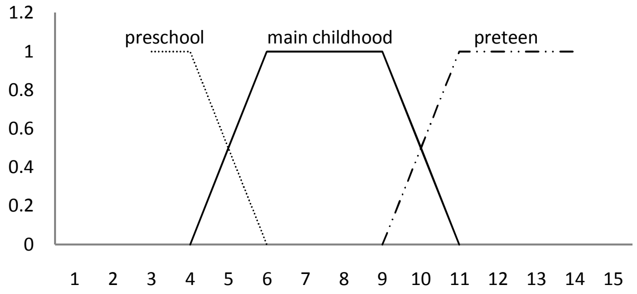

Figure 2. Fuzzy sets representing the different childhood periods of young web users.

3. Pre-Ranking Document Valorization

After running a classical query/document process, we get the score relevance of each document to the query. The pre-ranking document valorization aims to increase score relevance of relevant documents depending on their particularities. It's much easier and time-saver for the system to handle only relevant documents and not all the collection documents. For that, we choose to not run the document valorization while the querying process but after it. We define three types of document valorization: the "multimedia richer" document valorization, the "Nearest personal information" document valorization and "same-topic" document valorization. All of the three types of the pre-ranking document valorization are submitted to the application of fuzzy rules.

3.1. "Multimedia Richer" Document Valorization

Given that younger user are interested more in multimedia objects (images, video sequences …) while the information retrieval process. The "multimedia richer" document valorization aims to increase the relevance score of relevant document including more multimedia objects. This fact is submitted to fuzzy rules aiming to decide the value added *V* to score relevance proportionally to the user age *UA* and the Number of Multimedia Objects in the Document *NMOD*.

- If UA is in preschool period and NMOD is high then V is high.
- If UA is in main childhood period and NMOD is high then V is medium.
- If UA is in preteen period and NMOD is high or medium then V is low.

The Number of Multimedia Objects in the Document *NMOD* is an input set represented by triangular membership functions which ranges from Min to Max (see figure 3). Min and Max are variables representing respectively the minimum number of multimedia objects included in a document in the collection and the maximum number of multimedia objects included in a document in the collection.

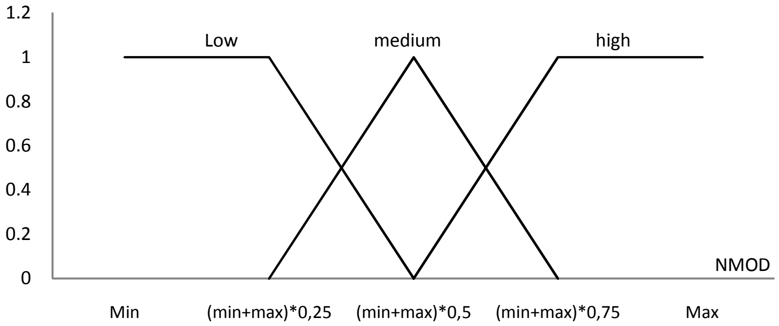

Figure 3. Fuzzy sets representing the variation range of NMOD.

3.2. "Nearest Personal Information" Document Valorization

As we mention before, more than 70% of surveyed children want to see in returned results information dealing with their own particulars (own country, own avenue, own school, friends …). The idea is to familiarize the information searching concept to children through the maximum coverage of their own particular in the information searching process. Referring to this observation, we suppose that results will be considered relevant if they include some personal data. The nearest personal information document valorization exploits the user age *UA* and the Number of Personal information Items found in a Document *NPID* as numerical inputs of the fuzzification operation submitted to the fuzzy rules. A personal information item found in a document may be a piece of text, a picture, or any multimedia object dealing with the current user particularities. The fuzzy rules listed below make decision about the value *V* added to a document relevance score in order to valorize documents dealing with user personal information:

- If UA is in main childhood period and NPID is high then V is medium.
- If UA in preteen period and NPID is high then V is high.
- If UA in preteen period and NPID is medium or low then V is low.

As the *NMOD* variable, The *NPID* is an input set represented by triangular membership functions which ranges from 0 to Max (see figure 4). Max is a variable representing the maximum number of personal information items concerning a user found in a document of the collection and 0 means that a document didn't include any information items concerning the current user.

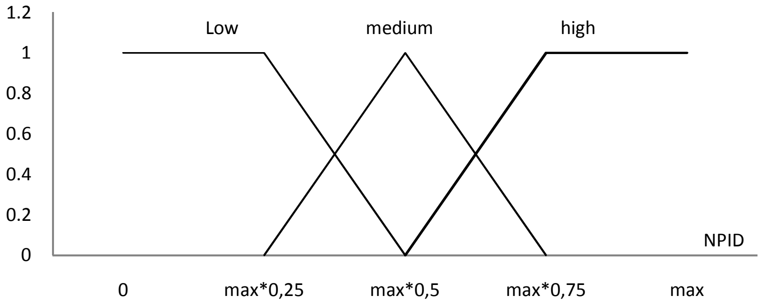

Figure 4. Fuzzy sets representing the variation range of NPID.

3.3. "Same-Topic" Document Valorization

A "Same-topic" document valorization aims to increase relevance score of document referring to the topics mentioned therein. The Figure 5 gives a structural view of this type of document valorization. The meta-document is introduced in [9] and it is able to annotate multimedia objects as well as web documents in a way that ensures its reusability. The querying process matches the user query with the meta-documents in order to identify the score relevance of the document to the query. We define the "topic cloud" as groups of weighted terms concerning a given topic. Simply, we collect potential terms representing a given topic to construct a topic cloud. The terms' weights express the ability of each term to represent the topic. After running a usual querying process matching the query and the meta-documents, we get the relevance score for each annotated resource or document. At this point, the topic clouds are used to enhance ranking results in the benefit of relevant documents focusing mainly on query interests.

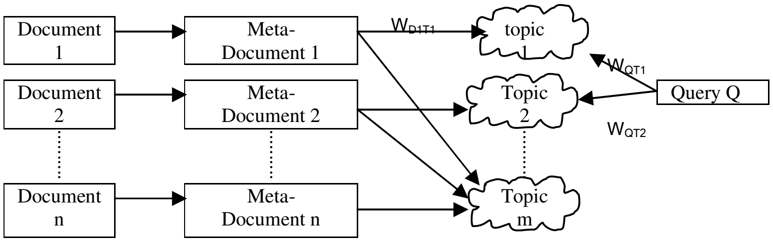

Figure 5. The "Same-topic" document valorization.

To run the "Same-topic" document valorization, first, we establish the meta-document/topic weighted links W_{DT}. $\mathrm{W_{DT}}$ expresses the potential topics mentioned by the document. To assign a weight W_{DT} to a meta-document/topic link, we simply sum the weights of topic terms existing in the meta-document. Then we establish query/topics weighted links which express the ability of each topic to represent the query. To assign a weight W_{QT} to a Query/Topic link, we use the classic similarity measure between two weighted terms vectors:

$$\mathrm{W_{QT}} = \mathrm{sim(Q, Ti)} = \frac{\sum_{j=1}^{t} \mathrm{W_{qj}} * \mathrm{W_{t_{ij}}}}{\sqrt{\sum_{j=1}^{t} (\mathrm{W_{qj}})^2 * \sum_{j=1}^{t} (\mathrm{W_{t_{ij}}})^2}} \quad (1)$$

The next step of the "Same topic" document valorization is to calculate for each document his topic similarity relative to the query in order to increase or decrease its relevance score in terms of the value of the topic similarity. The topic similarity TS is calculated as follow:

$$\mathrm{TS\ (Q, D)} = \sum_{i=1}^{k} \left| \mathrm{W_{QT_i}} - \mathrm{W_{DT_i}} \right| \quad (2)$$

The main goal of a "Same topic" document valorization is to increase relevance of documents focusing on the same query topics and valorize "mono-topic" documents to users at an early age in order to facilitate its comprehension. The *TS (Q, D)* value is optimal when its value is minimal; this means that the query and the document are focusing on the same topics with approached values. Contrariwise, if the *TS* is high, this means that the document deals with other topics in addition to the query topics. Finally, the increase value *V* affected to a document Relevance Score *RS* is based on the following fuzzy rules:

— If UA is in (pre-school or main childhood period) and TS is low then V is high
— If UA is in preteen period and TS is low or medium then V is medium

Remains to mention that this valorization method is inspired from our previous work [10], except that we exclude the rule that decrease relevance score of documents having a high TS value. This exclusion allows to not penalizing some document because of the diversity of topics included therein. Also, we include the UA variable in this valorization method instead of score relevance RS. The *TS* variable is an input set represented by triangular membership functions which ranges from 0 to Max (see figure 6). Maximum *TS* means that the current document deals with several topics in addition to the query topics and a null *TS* means that the document and the query are dealing with the same topics.

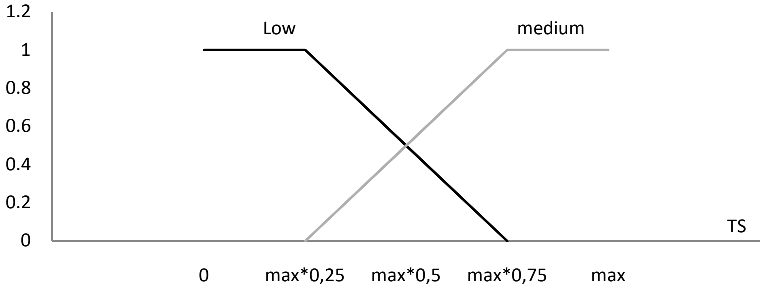

Figure 6. Fuzzy sets representing the variation range of TS.

After the application of valorization methods separately to a document D_i we sum the eventually deduced values to get the value V_i which is added to the document score relevance in order to get the new score relevance. After applying the valorization process to the relevant documents set found in the querying process, we pass finally to the ranking procedure. Table 1 summarizes the document valorization process.

Table 1. Recapitulation of the pre-ranking document valorization process.

Valorization method	Inputs variables		output
Multimedia Richer	NMOD	UA	v_1
Nearest Personal Information	NPID	UA	v_2
Same-Topic	TS	UA	v_3
Valorization process	Document D_i		$V_i=\sum_{k=1}^{3} v_k$

4. CONCLUSION

In this paper, we present three methods to valorize score relevance of some documents depending on their characteristics concerning the multimedia included objects, the user particulars and the topics mentioned therein. In this work, we use fuzzy rules to define in a flexible way the value added to the score relevance of valorized documents. Our framework is under development and it represents an information retrieval system dedicated for kids. We are working in the short-term on the identification of the relevant range and shape of the membership function representing the added value *V* usable on the three valorization methods.

REFERENCES

[1] Tim Berners-Lee, James Hendler and Ora Lassila the Semantic Web. Scientific American: Feature Article: The Semantic Web: May 2001

[2] W3C. Semantic Web http://www.w3.org/standards/semanticweb/

[3] Zadeh, L.A.: Fuzzy sets. Information and Control (1965) 338–353 http://www-bisc.cs.berkeley.edu/Zadeh-1965.pdf

[4] Lotfi A. Zadeh: Knowledge Representation in Fuzzy Logic. IEEE Trans. Knowl. Data Eng. 1(1): 89-100 (1989)

[5] Lotfi A. Zadeh: A Summary and Update of "Fuzzy Logic". GrC 2010: 42-44

[6] Manola, F. Miller, E., Beckett, D. and Herman, I. 2007. RDF Primer http://www.w3.org/2007/02/turtle/primer/

[7] RDF Working Group. Resource Description Framework (RDF) http://www.w3.org/RDF/

[8] Eric Prud'hommeaux, W3C. SPARQL Query Language for RDF W3C Recommendation 15 January 2008 http://www.w3.org/TR/rdf-sparql-query/

[9] Jedidi A. (July 2005). « modélisation générique de documents multimédia par des métadonnées : mécanismes d'annotation et d'interrogation » Thesis of « Université TOULOUSE III Paul Sabatier », France.

[10] Chkiwa, M., Jedidi, A., Gargouri, F. (October 2013). Handling Uncertainty in Semantic Information Retrieval Process. URSW 2013: 29-33, Sydney Australia