Possibilistic Networks for Aggregated Search in XML Documents

Fatma Z. Bessai-Mechmache[1], **Zaia Alimazighi**[2]

[1]Research Centre on Scientific and Technical Information, CERIST, Algiers, Algeria
zbessai@cerist.dz
[2]University of Science and Technology, USTHB, LSI, Algiers, Algeria
alimazighi@wissal.dz

ABSTRACT
In this paper, we are interested in aggregated search in conten-based XML information retrieval. Our goal is to revisit the granularity of the unit to be returned. More precisely, instead of returning the whole document or a list of disjoint elements of a document, we attempt to build the best element aggregation (set of non redundant elements) which is likely to be relevant to a query. For this, we present a model for XML information retrieval, based on possibilistic networks. The network structure provides a natural representation of links between a document, its elements and its content, and allows an automatic selection of relevant and complementary elements. Experiments carried out on a sub-collection of INEX (INitiative for the Evaluation of XML retrieval), showed the effectiveness of the approach.
Keywords
XML Information Retrieval, possibilistic network, aggregated search, complementary elements.

1. INTRODUCTION

The main problem of content-based XML information retrieval is how to select the unit of information that better answers the user's query composed of only key words (content only) [9] [12].
Most of XML Information Retrieval (IR) approaches [18] [13] [10] [11] [14] consider that the returned units are a list of disjoint elements (subtrees of XML documents). We assume that relevant unit is not necessarily a unique adjoining elements or a document it could also be any aggregation of elements of that document. Let us consider a document with the following structure *(document (title)(chappter1(section1)(section2))chapter2(...))* if the relevant information are located in the "title" and "section1", the majority of XML IR systems will return the whole document as the relevant unit, in our case we consider that, the only unit to be returned is an aggregate (element set) formed by both elements : "title" & "section1". To achieve this objective, we propose a model enabling to automatically select aggregation of non redundant elements of the document that better answer the user's need formulated through a list of key words. The model we propose finds its theoretical bases in the possibilistic networks. The network structure provides a natural manner to represent the links between, a document, its elements and its content. As for the possibilitic theory, it makes it possible to quantify in a qualitative and quantitative way the various subjacent links. it allows to express the fact that a term is certainly or possibly relevant with respect to an element and/or a document and to measure at which point an element (or a set of elements) can necessarily or possibly answer the user's query.

This paper is organized in the following way. Section 2 presents a brief state of the art on aggregation search. Section 3 gives a brief definition of the possibilistic theory. Section 4 is devoted to the description of the model which we propose. We show, in section 5 an example illustrating this model. Section 6 gives the evaluation results and showed the effectiveness of the model. Section 7 concludes the paper.

2. STATE OF THE ART

The aim of the aggregated search is to assemble information from diverse sources to construct responses including all information relevant to the query.
The issue of aggregation of elements from a collection of XML documents is not addressed in the literature. Indeed, the proposed approaches that address this issue are limited to Web documents [6] [1]. However, few Information retrieval systems begin to aggregate the results of a query on XML documents as summaries. For example, eXtract [8] is an information retrieval system that generates results as XML fragments. An XML fragment is qualified like result if it answers four features: Autonomous (understanding by the user), distinct (different from the other fragments), representative (of the themes of the query) and succinct. XCLUSTERs [15] is a model of representation of XML abstracts. It regroups some XML elements and uses a small space to store the data. The objective is to provide significant excerpts so users can easily evaluate the relevance of query results.
The approach we propose in this paper is located to junction between the research of the relevant elements and their regrouping (aggregation) in a same result. Our approach is based on the possibilistic theory [19] [7] [4] and more particularly the possibilistic networks [2] [3]. These networks offer a simple and natural model for representing the hierarchical structure of XML documents and to handle uncertain inherent in information retrieval. One finds this uncertainty in, the concept of relevance of a document with respect to a query, the degree of representativeness of a term in a document or part of documents and the identification of the relevant part answering the query. Within this framework, in order to identify the relevant part which answers the query, contrary to the approaches suggested in the literature, which select as we have seen, the sub-tree likely to be relevant, our approach allows to identify and to select, in a natural way, the element or an aggregation of non redundant elements of XML document likely to answer the query.

Besides the above-stated points, the theoretical framework which supports our proposals, in fact the possibilistic networks

clearly differentiate us from the settings used in the previous approaches.

3. THE POSSIBILISTIC THEORY

The possibilistic logic [19] allows flexibility in the available data processing. It enables to model and quantify the relevance of a document considering a query through two measurements: the necessity and the possibility. The necessarily relevant elements are those which must appear in top of the list of the selected elements and must allow certain system efficiency. The possibly relevant elements are those that would eventually answer the user query. They appear in the list of the selected elements classified following the necessarily relevant elements or failing this (if the system does not find any) they are regarded as a plausible answer.

Possibility distribution:

A possibility distribution π is a mapping from X to [0, 1]. π (x) evaluates to what extent x is the actual value of some variable to which π is attached. π (x) =1 means that it is completely possible that x is the real world (or that x is completely fulfilling), 1> π (x) >0 means that x is somewhat possible (or fulfilling), and finally π (x) =0 means that x is certainly not the real world (or is completely unsatisfactory). An event is said 'no possible' does not only mean that the opposite event is possible. It actually means that it is certain. Two dual measures are used: the possibility measure $\Pi(A)$ and the necessity measure N(A). The possibility of an event A, noted $\Pi(A)$ is obtained by $\Pi(A) = \max_{x \in A} \pi(x)$ and describes the most normal situation in which A is true.

The necessity $N(A) = \min_{x \notin A} 1 - \pi(x) = 1 - \Pi(\neg A)$ of an event A reflects the most normal situation in which A is false.

Possibilistic conditioning:

In the possibilistic setting, the possibilistic conditioning consists in modifying our initial knowledge, encoded by the possibility distribution π by the arrival of new fully certain piece of information e. Let us denote Φ = [e] the set of models of e. The initial distribution π is then replaced by another one denoted by $\pi' = \pi(\bullet/\Phi)$. Assuming that $\Phi \neq \emptyset$ and that $\Pi(\Phi) > 0$, the natural postulates for possibilistic conditioning are:

$\pi(w /p \Phi) = \pi(w)/ \Pi(\Phi)$ if $w \in \Phi$ (1)

and 0 otherwise

Where /p is the product-based conditioning.

Product-based possibilistic network:

A product-based possibilistic network over a set of variables $V = \{A_1, A_2, \ldots, A_N\}$ is a possibilistic graph where conditionals are defined using product-based conditioning.

The possibility distribution of the product-based possibilistic network, denoted by Π_P, is obtained by the following rule of chaining [2]:

$\Pi_P (A_1,..., A_N) = PROD_{i=1..N} \Pi (A_i/PAR_{Ai})$ (2)

where PROD is the product operator

4. POSSIBILISTIC MODEL FOR XML INFORMATION RETRIEVAL

4.1 Model Architecture

The architecture of the model we propose is illustrated in figure (1). The graph allows to represent the documents nodes, index terms, nodes (elements of an XML document). The links between the nodes allow representing the relations of dependences between the various nodes.

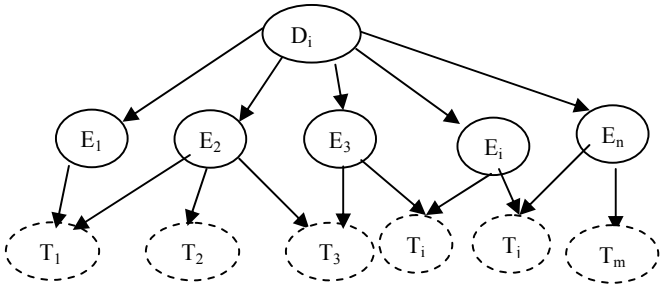

Figure1. Architecture of the model

The documents nodes represent the documents of the collection. Each document node Di, represents a binary random variable taking values in the set dom $(D_i) = \{d_i, \neg d_i\}$, where the value $D_i = d_i$ (resp. $\neg d_i$) represents "the document D_i is relevant for a given query (resp. non-relevant).

Nodes $E_1, E_2, \ldots, E_n$, represent the elements of document D_i. Each node E_i, represents a binary random variable taking values in the set dom $(E_i) = \{e_i, \neg e_i\}$. The value $E_i = e_i$ (rep. $\neg e_i$) means that the element 'E_i' is relevant for the query (resp. non-relevant).

Nodes $T_1, T_2, \ldots, T_m$ are the terms nodes. Each term node Ti represents a binary random variable taking values in the set dom $(T_i) = \{t_i, \neg t_i,\}$ where the value $T_i = t_i$ (resp. $\neg t_i$) means that term 'T_i' is representative of the father node to which it is attached (resp. non-representative of the father node to which it is connected). It should be noticed that a term is connected to the node that includes it as well as to all its ancestors.

The passage of the document towards the representation in the form of possibilistic network is done in a simple way. All the nodes (elements) represent the level of variables E_i. The values that will be assigned with the arcs of dependences between term nodes – element nodes and element nodes – document node depend on the sense which one gives to these links.

Each structural variable E_i, $E_i \in E = \{E_1, E_2, \ldots, E_n\}$, depends directly on its parent node which is the node root D_i in the possibilistic network of the document. Each variable of contents T_i, $T_i \in T = \{T_1, \ldots, T_m\}$ depends only on its structural variable (structural element or tag). It should be also noticed that the representation considers only one document. In fact one considers that the documents are independent from each other thus one can only consider the sub-network representing the document that is processed.

We note by T(E) (resp. T(Q)) the set of the index terms of the elements of the document (resp. of the query).

The arcs are oriented and are of two types:

- Term-element links. These links connect each term node $T_i \in T(E)$ to each node E_i where it appears.
- Elements-document links. These links connect each element node of the set E to the document that includes it, in our case D_i.

In sections 4.2.1 and 4.2.2, we will discuss the interpretation we give these various links and the way we quantify them.

4.2 Query Evaluation

As we underlined previously, we model the relevance according to two dimensions: the necessity and the possibility of relevance. Our model must be able to infer propositions of the type:

- "the document d_i is relevant for the query Q" is possible to a certain degree or not, quantified by $\Pi(Q/d_i)$.

- "the document d_i is relevant for the query Q" is certain or not, quantified by $N(Q/d_i)$.

The first type of proposition allows to eliminate the non-relevant documents (and elements), i.e. those that have a weak possibility. The second proposition focuses the attention on those that seem very relevant.

For the model presented here, we will adopt the following assumptions:

Assumption1: A document has as much possibility to be relevant than non-relevant for a given user, either $\Pi(d_i) = \Pi(\neg d_i) = 1$, $\forall i$.

Assumption2: The query is composed of a simple list of key words $Q= \{t_1,t_2, \dots ,t_n\}$. The relative importance between terms in the query is ignored.

According to the definitions of the possibilistic theory, the quantities $\Pi(Q/d_i)$ and $N(Q/d_i)$ are calculated like follows :

$$\Pi(Q/d_i) = \max_{\forall \theta^e \in \theta^E} \left(\text{Prod}_{E_j \in \theta^e} \left(\text{Prod}_{T_i \in T(E) \wedge T(Q)} (\Pi(t_i / \theta_j^e)) \right) * \text{Prod}_{E_j \in \theta^e} (\Pi(\theta_j^e / d_i)) * \Pi(d_i) \right)$$

(3)

Where:

- Prod: means product (we used this symbol instead of $\prod$ not to confuse it with the symbol designating the possibility).
- $t_i \in T(E) \wedge T(Q)$: represents the terms of the queries which index the elements of the XML document.
- θ^e : set of non redundant elements
- θ_j^e : represents the value of E_j in the aggregation θ^e (example: the value of E_1 in the aggregation $\{e_1 \wedge e_2\}$ is e_1).

We recall, as we underlined previously, that the selection of the relevant parts (units of information) is inherent with the model. Indeed, the formula (3) calculates the relevance by considering all possible aggregations (combinations) of elements. The factor θ^e gives possible values of elements. The aggregation of elements that will be selected will be the one that includes obligatorily the terms of the query and presents the best relevance (maximum relevance) in terms of necessity and/or possibility.
As it was mentioned in the introduction, our model is able tri select the best aggregation of elements that are likely to be relevant to the query. This aggregation is the aggregation that maximises the necessity if it exists or the possibility. It obtained by:

$$\theta^* = \arg\max_{\forall \theta^e \in \theta^E} \Pi(Q/d_i)$$

The various degrees Π and N between the nodes of the network are calculated as follows:

4.2.1 Possibility distribution $\Pi(t_i/e_j)$

In Information Retrieval, the terms used to represent the content of a document, are weighted in order to better characterize the content of this document. The same principle is used in XML retrieval. The weights are generally calculated by using term frequency (tf) within a document or inverse document frequency (Idf) in the collection.

In information retrieval, it has been shown [16] [17] that the performances of the system can be improved if one represents an element by considering its own content and the contents of its sons elements. In our model, we distinguish the terms possibly representative of the elements of the document and those necessarily representative of these elements (terms that are sufficient to characterize the elements). With this intention, the possibility of relevance of a term (t_i) to represent an element (e_j), noted $\Pi(t_i/e_j)$, is calculated like follows:

$$\Pi(t_i / e_j) = tf_{ij} / \max_{\forall t_k \in e_j} (tf_{kj})$$

Where: tf_{ij} represents the frequency of the term 't_i' in the element 'e_j'.

A term having a degree of possibility 0 means that the term is not representative of the element. If the degree of possibility is strictly higher than 0, then the term is possibly representative of the element. If it appears with a maximum degree of possibility, then it is considered as the best potential candidate for the representation and thus the restitution of the element.

Let us note that: $\max(\Pi(t_i / e_j)) = 1$, $\exists t_i \in e_j$

In an XML document, a necessarily representative term of an element is a term that contributes to its restitution in response to a query. This term is called discriminative term and it is a term that frequently appears in few elements of XML document [5]. The factor commonly used in IR to quantify the discriminative power of a term is idf (ief in XML IR). Therefore, a degree of necessary relevance, β_{ij}, of the term t_i to represent the element e_j, will be defined by:

$$N(t_i \rightarrow e_j) \geq \beta_{ij} = \mu(tf_{ij} * ief_{ij}) * idf = \mu\left(tf_{ij} * \log\left(\frac{Ne}{ne_i + 1}\right) * \log\left(\frac{N}{n_i + 1}\right)\right)$$

With:

- N and Ne: respectively the number of documents and elements in the collection.
- n_i et ne_i : respectively the number of documents and the number of elements containing the term t_i.
- μ : a function of normalization. A simple manner to normalize is to divide by the maximal value of the factor.
- tfij: represents the frequency of the term 't_i' in the element 'e_j'.
- ief_{ij}: represents the inverse frequency of the element 'e_j' for the term 't_i'.
- idf: represents the inverse frequency of the document.

It should be noticed that this formula has been chosen according some experiments that were undertaken by Sauvagnat [17].
This degree of necessary relevance allows limiting the possibility that the term is compatible with the rejection of the element by:
$\Pi (t_i / \neg e_j) \leq 1 - \beta_{ij}$ (this is deduced by definition in the possibilistic theory)
We summarises the distribution of possibility on the Cartesian product $\{e_j, \neg e_j\}$ x $\{t_i, \neg t_i\}$ by the following table:

Table1. Possibility distribution on the set of terms T

Π	e_j	$\neg e_j$
t_i	$tf_{ij} / max(tf_{kj}), (\forall t_k \in e_j)$	$1 - \beta_{ij}$
$\neg t_i$	1	1

4.2.2 Possibility distribution $\Pi(e_j/d_i)$

The arc document node - element node (or arc root-element) indicates the interest to propagate information from an element towards the document node (root). The nodes appearing close to the root (of a tree) are carrying more information for this node root than those located lower in the tree [17]. Thus it seems intuitive that more an element is far from the root more it is less relevant. We model this intuition by the use in the function of propagation of the parameter *dist(root, e)*, that represents the distance between the root node and one of its descendant nodes (elements) 'e' in the hierarchical tree of the document, i.e. the number of arcs separating the two nodes.

The degree of possibility of propagation of relevance of an element (e_i) towards the document node (d_i) is defined by $\Pi (e_j / d_i)$ and is quantified as follows.

$$\Pi (e_j / d_i) = \alpha^{dist(di, ej)-1}$$

With:
- $dist(d_i, e_j)$ the distance from the element e_j to the root d_i in accordance with the hierarchical structure of the document.
- $\alpha \in]0..1]$ is a parameter allowing to quantify the importance of the distance separating the elements nodes (structural elements of the document) to the root of the document.

Concerning the necessity to propagate, in an intuitive manner, one can think that the designer of a document uses the nodes of small size to emerge important information. These nodes can thus give precious indications on the relevance of their ancestors' nodes. A title node in a section for example allows locating with precision the subject of its ancestor node section. It is thus necessary to propagate the signal calculated on the level of the node towards the root node. To answer this intuition, we propose to calculate the necessity of propagation of relevance of an element e_j towards the root node d_i, noted $N(e_j \rightarrow d_i)$, as follows:

$$N(e_j \rightarrow d_i) = 1 - \frac{le_j}{dl}$$

le_j is the size of the element node ej and dl the size of a document (in number of terms). According to the formula, the more a term is of small size, the bigger is the necessity to propagate it.

Therefore, $\Pi (e_j/ \neg d_i) = le_j/dl$

We recapitulate the distribution of possibility definite on the Cartesian product $\{d_i, \neg d_i\}$ x $\{e_j, \neg e_j\}$ by the following table:

Table 2. Possibility distribution on the set of elements E

Π	d_i	$\neg d_i$
e_j	$\alpha^{dist(d_i, e_j)-1}$	le_j/dl
$\neg e_j$	1	1

5. Illustrative example

An example of XML document (an extract of a document) related to a book will be used to illustrate our talk. The XML document and its possibilistic network are presented as follows:

```
<Book>
  <Title > Information Retrieval </Title >
  <Abstarct> In front of the increasing mass of information …</Abstract>
….
  <Chapter>
     <Title chapter> Indexing </title chapter>
     <Paragraph> The indexing is the process intended to represent by the elements of a documentary or natural language of … </Paragraph>
  </Chapter>
</Book>
```

The possibilistic network associated with XML document `Book' is as follows:

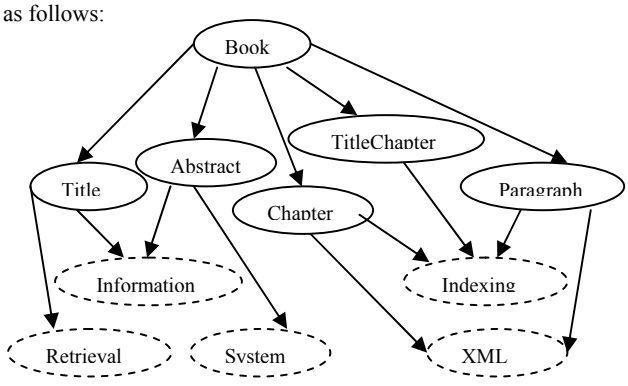

Figure 2. Possibilistic network of the XML document 'Book'

For this example, the set of the elements E= $\{e_1$=Title, e_2=Abstract, e_3=Chapter, e_4=Titlechapter, e_5=Paragraph$\}$. The set of the indexing terms of the elements, calculated while using the content of each element, along with its child elements in the document, such as T(E) = $\{t_1$=Retrieval, t_2=Information,

t_3=System, t_4=Indexing, t_5=XML}. We consider only some terms not to congest the example.
The matrix containing the values of the arcs element node- term node of the possibilistic network of the document "Book" is given in table 3. We recall that a term is connected to the node that includes it as well as to all the ancestors of this node.

Table3. Distribution of possibility Π (t_i/e_j)

Π (t_i/e_j)	t_1	t_2	t_3	t_4	t_5
e_1 ¬ e_1	1 0	1 0	0 1	0 1	0 1
e_2 ¬ e_2	1/2 0.50	1 0	1 0	1/4 0.88	0 1
e_3 ¬ e_3	0 1	0 1	0 1	0.70 0.10	0.50 0.20
e_4 ¬ e_4	0 1	0 1	0 1	1 0	0 1
e_5 ¬ e_5	0 1	0 1	0 1	0.88 0.05	1 0

The matrix containing the values of the arcs root- element node of the possibilistic network of the document `Book' is given in table 4 (we take α = 0,6 and dl=100).

Table4. Distribution of possibility Π (e_j/d_i)

	Π (e_j/d_i) (d_i=book)	Π (e_j/d_i) (d_i= ¬book)
e_1	1	0.02
¬e_1	1	1
e_2	1	0.1
¬e_2	1	1
e_3	1	1
¬e_3	1	1
e_4	0.6	0.01
¬e_4	1	1
e_5	0.6	1
¬e_5	1	1

When the query is put, a process of propagation is started through the network modifying the values of possibilities a priori. In this model the formula of propagation used is the formula (3).

Let's take a query Q composed of the keywords "Retrieval" and "Information", Q = {Retrieval, Information}.
According to the assumption 1, Π (d_i) = Π(¬ d_i) = 1, ∀i.
Given the query Q, the propagation process (formula (3)) considers only the aggregates of set E that include the query terms t_1 = 'Retrieval' and t_2= 'Information'. In fact only the elements e_1='Title' and e_2='Abstract' will be considered. The aggregations that it is thus necessary considered are: {$e_1 \wedge e_2$, $e_1 \wedge \neg e_2$, $\neg e_1 \wedge e_2$, $\neg e_1 \wedge \neg e_2$}. We calculate then:

For d_i = book:

a_1 = Π(t_1/e_1).Π (t_2/e_1). Π (t_2/e_2). Π (e_1/book). Π (e_2/ book) = 1 * 1*1 * 1 * 1= 1

a_2 = Π(t_1/e_1).).Π (t_2/e_1).Π (t_2/¬e_2). Π (e_1/book). Π (¬e_2/book) = 1* 1*0 * 1 * 1= 0

a_5 = Π (t_1/¬e_1).).Π (t_2/¬e_1).Π (t_2/e_2). Π (¬e_1/book). Π (e_2/book) = 0 * 0*1 * 1 * 1 = 0

a_4 = Π(t_1/¬e_1). Π(t_2/¬e_1). Π (t_2/¬e_2). Π (¬e_1/book). Π (¬e_2/ book) = 0 * 0 * 0 * 1 * 1 = 0

According to the formula (3):
Π(Q/book) = max (a_1, a_2, a_3, a_4) = 1 = a_1

For ¬ d_i = ¬ book:

a_5 = Π (t_1/e_1). Π (t_2/e_1). Π (t_2/e_2). Π(e_1/¬book). Π(e_2/¬book) = 1 * 1 * 1 * 0.02 * 0.1 = 0.002

a_6 = Π (t_1/e_1). Π (t_2/e_1). Π (t_2/¬e_2). Π(e_1/¬book). Π(¬e_2/¬book) = 1 * 1* 0 * 0.02 * 0.1 = 0

a_7 = Π (t_1/¬e_1). Π (t_2/¬e_1).Π (t_2/e_2). Π(¬e_1/¬book).Π(e_2/¬book) = 0 * 0 * 1 * 1 * 1 * 0.1 = 0

a_8 = Π (t_1/¬e_1).Π(t_2/¬e_1).Π (t_2/¬e_2).Π(¬e_1/¬book).Π(¬e_2/¬book) = 0 * 0* 0 * 1 * 1 = 0

According to the formula (3):
Π(Q/¬ book/) = max (a_5, a_6, a_7, a_8) = 0.002 = a_5
To calculate the necessity N(Q/book) =1- Π(Q/¬ book) = 1- 0.002 = 0.998
To calculate the necessity N(Q/¬ book) =1- ∏(Q/book/) = 1- 1 = 0
The preferred documents are those that have a value N (Q/d_i) high among those that have a value Π (Q/d_i) high too. If N (Q/d_i) = 0, the restored documents are (unwarranted of total adequacy) those that have a value Π (Q/d_i) high. Therefore, for the query Q = {Retrieval, Information}, it is the aggregation "a_1" (title, abstract) that will be turned to the user as answer to his query.

6. EXPERIMENTS AND RESULTS

The main objective of this assessment is to show the importance of aggregated search in XML documents.
All studies performed to assess aggregated search were based on usage studies. To evaluate our model a prototype was developed.
Our experiments are conducted on a sample about 3000 XML documents of the INEX'2005 collection, a set of 20 queries from the same collection and 30 users. Each user evaluates six queries. There will be 180 users' judgments to analyze for the set of the twenty queries.
The following histogram gives the judgments of users by query regarding the aggregate relevance.

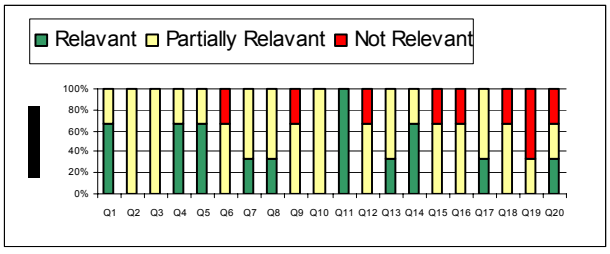

Figure 3: Distribution of aggregation relevance results

The experimental evaluation shows that aggregated search has big contribution for XML information retrieval. Indeed, the aggregate gathers non-redundant elements (parts of XML document). These elements can be semantically complementary and in this case the aggregate allows improving the interpretation of results, guides the user to the relevant elements of XML document, faster and also reduces the efforts the user must provide to locate information searched for. However, in some cases elements of the aggregate may be non complementary that means not semantically related with respect to information need expressed by user's query. This sort of aggregation is very useful because it allows a very fine distinction of the different thematic expressed in the user's query when his need in information is generic. It also helps inform the user about various information of the corpus related to his information need thus help him, if necessary, to reformulate his query.

7. CONCLUSION

This paper presents a new approach for XML information retrieval based on possibilitic networks. This model provides a formal setting to aggregate non-redundant and complementary elements into the same unit.

Our aggregated search model for XML documents is characterized by the following main points:

- Directs the user more quickly toward the relevant elements of XML document.
- Informs the user on various information of the corpus related to his information need.
- Helps to query reformulation.

Thus, it seems very interesting to think of defining assessment tools that evaluate the aggregated search systems in order to compare the different aggregation models.

8. REFERENCES

[1] Agrawal R., Gollapudi S, Halverson A. Diversifying Search Results. ACM Int. Conference on WSDM, 2009.

[2] Ben Amor N. *Qualitative Possibilistic Graphical Models : From Independence to Propagation Algorithms*, Thèse pour l'obtention du titre de Docteur en Gestion, université de Tunis, 2002.

[3] Benferhat, S., Dubois, D., Garcia, L., Prade, H. Possibilistic logic bases and possibilistic graphs. In Proc. of the 15th Conference on Uncertainty in Artificial Intelligence, 57-64, 1999.

[4] Borgelt, C., Gebhardt, J. and Kruse, R. Possibilistic graphical models. Computational Intelligence in Data Mining, CISM Courses and Lectures 408, Springer, Wien, 51-68, 2000.

[5] Brini, A., Boughanem, M., Dubois, D. A Model for Information Retrieval Based on Possibilistic Networks. SPIRE'05, Buenos Aires, 2005. LNCS, Springer Verlag, p. 271-282.

[6] Clarke C. L., Kolla M., Cormack G. V., Vechtomova O. Novelty and diversity in information retrieval evaluation. SIGIR'08, p. 659-666, 2008.

[7] Dubois, D. and Prade, H.Possibility Theory. Plenum, 1988.

[8] Huang Y., Liu Z., Chen y. Query biased snippet generation in XML search. ACM SIGMOD, p. 315-326, 2008.

[9] Kamps, J., Marx, M., De Rijke M., Sigurbjörnsson B. XML Retrieval: What to retrieve? ACM SIGIR Conference on Research and Development in Information Retrieval, p.409-410, 2003.

[10] Lalmas, M. Dempster-Shafer's Theory of Evidence Applied to Structured Documents: Modelling Uncertainty. In Proceedings of the 20th Annual International ACM SIGIR, pages 110–118, Philadelphia, PA, USA. ACM. (1997).

[11] Lalmas, M., Vannoorenberghe, P. Indexation et recherche de documents XML par les fonctions de croyance. CORIA'2004, pp 143-160 (2004).

[12] Fuhr, N., Lalmas, M., Malik, S., Szlavik, Z. Advances in XML Information Retrieval: INEX 2004, Dagstuhl Castle, Germany, December 6-8, 2004.

[13] Ogilvie, P. and Callan, J. Using language models for flat text queries in XML retrieval. In Proceedings of INEX 2003 Workshop, Dagstuhl, Germany, pages 12–18, December 2003.

[14] Piwowarski, B., Faure, G.E., Gallinari, P. Bayesian Networks and INEX. In INEX 2002 Workshop Proceedings, p. 149-153, Germany, 2002.

[15] Polyzotis N., Garofalakis M. N. XCluster Synopses for Structured XML Content. ICDE, p. 63, 2006.

[16] Rölleke, T., Lalmas, M., Kazai, G., Ruthven, I., Quicker, S. The accessibility Dimension for Structured Document Retreival. BCS-IRSG European Conference on Information Retrieval (ECIR), Glasgow, Mars 2002.

[17] Sauvagnat, K. *Modèle flexible pour la Recherche d'Information dans des corpus de documents semi-structurés*. Thèse de Doctorat de l'Université Paul Sabatier, Juillet 2005.

[18] Sigurbjornsson, B., Kamps, J. and de Rijke, M. An element-based approach to XML retrieval. INEX 2003 workshop, Dagstuhl, Germany, December 2003.

[19] Zadeh, L. A. Fuzzy Sets as a Basis for a theory of possibility. In Fuzzy Sets and Systems, 1:3-28, 1978.