Aggregated Search in XML Documents

Fatma Zohra Bessai-Mechmache
Research Centre on Scientific and Technical Information, CERIST, Algiers, Algeria
zbessai@cerist.dz

Zaia Alimazighi
University of Science and Technology, USTHB, LSI, Algiers, Algeria
alimazighi@wissal.dz

***Abstract*—In this paper, we are interested in aggregated search in structured XML documents. We present a structured information retrieval model based on possibilistic networks. Relations terms-elements and elements-document are modeled through possibility and necessity. In this model, the user's query starts a process of propagation to recover the elements. Thus, instead of retrieving a list of elements that are likely to answer partially the user's query, our objective is to build a virtual elements that contain relevant and non-redundant elements, that are likely to answer better the query that elements taken separately. Indeed, the possibilistic network structure provides a natural representation of links between a document, its elements and its content, and allows an automatic selection of relevant and non-redundant elements. We evaluated our approach using a sub-collection of INEX (INitiative for the Evaluation of XML retrieval) and presented some results for evaluating the impact of the aggregation approach.**

***Index Terms*— XML Information Retrieval, possibilistic networks, aggregated search, redundancy.**

I. 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 [13] [9].

Most of XML Information Retrieval (IR) approaches [23] [17] [15] [16] [18] 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", most 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 as follows: 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 shows the effectiveness of the model. Section 7 concludes the paper.

II. 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. This comes in contrast with the common search paradigm, where users are provided with a list of information sources, which they have to examine in turn to find relevant content.

It is well known that, in the context of Web search, users typically access a small number of documents [12]. A study on users Web [24] showed that the percentage of users who consult fewer documents (Web pages) per query increases with time. For example, from 1997 to 2001, the users' percentage looking at a document by query is passed from 28.6% to 50.5%. This percentage increased further to over 70% after 2001 [25]. It gives to think that for a list of documents, is mainly confined to documents in the first, second and sometimes (at most) third rank. The study reported in [11] showed that on 10 documents displayed, 60% of users have looked less than 5 documents and nearly 30% have read a single document.

The aggregated search allows to bring solutions to this problem. Indeed, its aim is to integrate other types of documents (web documents, images, videos, maps, news, books) in the results page. Example of search engines that begin to make aggregation, we find Google Universal Search, Yahoo! alpha, etc. Users have access to various document types. This can be beneficial for certain

doi:10.4304/jetwi.4.2.181-188

queries, such as "trip to Finland" can return maps, blogs, weather, etc.

Another technique that can be used to improve the search results page is the clustering. However, it is not enough simply to return the clusters. It is important to provide users an overview of the contents of the documents forming a cluster [25]. A common approach to provide such an overview is a summary of all documents of the cluster ('multi documents summary'). Examples of systems based on this technique, we find NewInEssence [20], QCS [8], etc.

The issue of elements aggregation 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 [10] is an information retrieval system that generates results as XML fragments. An XML fragment is considered a result if it answers four features: Autonomous (understood by the user), distinct (different from the other fragments), representative (of the themes of the query) and succinct. XCLUSTERs [19] is a model of representation of XML abstracts. It includs some XML elements and uses a small space to store data. The objective is to provide significant excerpts for users to 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 possibilistic theory [26] [7] [4] and more particularly on possibilistic networks [2] [3]. These networks offer a simple and natural model for representing the hierarchical structure of XML documents and to handle uncertainty, inherent to information retrieval. We find this uncertainty in the concept of document relevance 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 that answers the query. Within this framework, to identify the relevant part that answers the query, unlike the approaches suggested in the literature, which select the sub-tree, likely to be relevant; our approach allows to identify and to select, in a natural way, an aggregation of non redundant elements of XML document that may answer the query.

Besides the points mentioned above, the theoretical framework, that supports our proposals, namely the possibilistic networks, clearly differentiate us from the settings used in previous approaches.

III. THE POSSIBILISTIC THEORY

The possibilistic logic [26] 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 that must appear in top of the list of the selected elements and must allow system efficiency. The possibly relevant elements are those that would eventually answer the user query.

A. Possibility Distribution

A possibility distribution, denoted by π, corresponds to a mapping from *X* to the scale [0, 1] encoding our knowledge on the real world.

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 Π(A) and the necessity measure N(A).

The possibility of an event A, denoted Π(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.

B. Product-based Conditioning

In the possibilistic setting, the product-based conditioning consists of modifying our initial knowledge, encoded by the possibility distribution π by the arrival of new fully certain piece of information e. Let us take Φ = [e] the set of models of e. The initial distribution π is then replaced by another one π′, such as π′ = π (•/Φ). Assuming that Φ ≠ Ø and that Π (Φ) > 0, the natural postulates for possibilistic conditioning are:

$$\pi(w /_p \Phi) = \pi(w)/ \Pi (\Phi) \quad \text{if } w \in \Phi \qquad (1)$$
$$\text{and} \quad 0 \quad \text{otherwise}$$

Where /p is the product-based conditioning.

C. Possibilistic Networks

A directed possibilistic network over a set of variables $V= \{V_1, V_2,\ldots,V_n\}$ is characterized by:

- A graphical component composed of a Directed Acyclic Graph (DAG). The DAG structure encodes a set of independence relations between variables.

- A numerical component consisting in a quantification of different links in the DAG using the conditional possibilisties of each node in the context of its parents. Such conditional distributions should respect the following normalization constraints for each variable V_i of the set V:

Let U_{Vi} the set of parents of V_i

If U_{Vi} = Ø (i.e. V_i is a root), then the a priori possibility relative to V_i should satisfy:

$$\max_{ai} \Pi(a_i) = 1, \ \forall \ a_i \in D_{Vi}$$

If $U_{V_i} \neq Ø$ (i.e. V_i is not a root), then the conditional distribution of V_i in the context of its parents should satisfy:

$$\max_{a_i} \Pi(a_i / Par_{V_i}) = 1, \forall a_i \in D_{V_i}$$

Where Par_{V_i} is the set of possible configurations (aggregations) of parents of V_i

Using the definition of conditioning based on the product operator. This leads to the following definition of a product-based possibilistic network:

D. Product-based Possibilistic Network

A product-based possibilistic network over a set of variables $V = \{A_1, A_2,…, A_N\}$ is a possibilistic graph where conditionals are defined using product-based conditioning (1).

Product-based possibilistic networks are appropriate for a numerical interpretation of the possibilistic scale.

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

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

Where: 'PROD' is the product operator.

IV. THE AGGREGATED SEARCH MODEL

A. Model Architecture

The architecture of the proposed model is illustrated in Fig. 1. The graph represents the document nodes, index terms and element nodes (elements of XML document). The links between the nodes allow representing the relations of dependences between the various nodes.

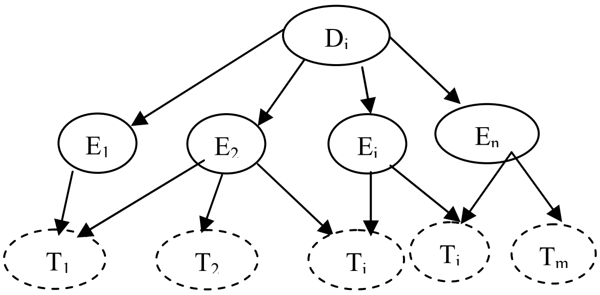

Figure 1 . Model Architecture.

Document nodes represent the collection documents. 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, …, 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, …, T_m are the term nodes. Each term node T_i 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 parent node to which it is attached (resp. non-representative of the parent 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 from the document to the representation in the form of possibilistic network is done in a simple way. All nodes (elements) represent the level of variables E_i. The values that will be assigned with the arcs of dependences between term-element nodes and element-document nodes depend on the sense which one gives to these links.

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

We will discuss interpretations we give these various links and the way we quantify them, in the following.

B. 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, 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:

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

Assumption 2: The query is composed of a simple list of key words Q= $\{t_1,t_2, \ldots ,t_n\}$. 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 :

$$\Pi(\mathrm{Q}/\mathrm{d}_{\mathrm{i}}) = \max_{\forall\theta^e\in\theta^E}(\mathrm{Pr}od(\mathop{\mathrm{Pr}od}_{E_j\in\theta^e}\;\mathop{\mathrm{Pr}od}_{T_i\in T(E)\wedge T(Q)}(\Pi(t_i/\theta_j^e)))*\mathop{\mathrm{Pr}od}_{E_j\in\theta^e}(\Pi(\theta_j^e/d_i))*\Pi(d_i)) \tag{3}$$

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, e_2) is e_1).

The selection of the relevant parts (units of information) is inherent with the model. Indeed, the (3) calculates the relevance by considering all possible aggregations (regrouping) 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 maximizes the necessity if it exists or the possibility. It obtained by:

$$\theta^* = \arg\max_{\forall\theta^e\in\theta^E}\Pi(\mathrm{Q}/\mathrm{d}_{\mathrm{i}}) \tag{4}$$

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

1) Possibility Distribution Π(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 [21] [22] that the performances of the system can be improved if one represents an element by considering its own content and the contents of its children nodes. 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:

$$\Pi\,(t_i\,/\,e_j) = \mathrm{tf}_{\mathrm{i\,j}}/\max_{\forall\,t_k\in e_j}(\mathrm{tf}_{\mathrm{kj}}) \tag{5}$$

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:

$$N(t_i \rightarrow e_j) \geq \beta_{ij} = \mu(tf_{ij} * ief_{ij}) * idf = \mu(tf_{ij} * \log(\frac{Ne}{ne_i + 1}) * \log(\frac{N}{n_i + 1})) \tag{6}$$

Where:

- N and N_e are respectively the number of documents and elements in the collection.
- n_i and ne_i are 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.
- tf_{ij} 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 (6) has been chosen according some experiments that were undertaken by Sauvagnat [22].

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 summarize the possibility distribution on the Cartesian product {e_j, ¬e_j} x {t_i, ¬ t_i} by the following table:

TABLE I. POSSIBILITY DISTRIBUTION ON THE SET OF TERMS T

Π	e_j	¬e_j
t_i	tf_{ij} / max(tf_{kj}), (∀t_k∈ e_j)	1- β_{ij}
¬t_i	1	1

2) Possibility Distribution Π(e_j /d_i)

The arc document-element (or arc root-element) indicates the interest to propagate information from an element towards the document node (root). The nodes, close to the root (of a tree), carry more information for the root than those located lower in the tree [22]. 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 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 Π (e_j / d_i) and is quantified as follows:

$$\Pi\ (e_j\ /\ d_i) = \alpha^{\ dist(di,\ ej)-1} \qquad (7)$$

Where:

- dist(d_i, e_j) is the distance from the element e_j to the root d_i in accordance with the hierarchical structure of the document.

- α ϵ]0..1] is a parameter allowing to quantify the importance of the distance separating the element 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, denoted $N(e_j \rightarrow d_i)$, as follows:

$$N(e_j \rightarrow d_i) = 1 - \frac{le_j}{dl} \qquad (8)$$

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

Therefore, Π (e_j/ ¬ d_i) = le_j/dl

We summarize the possibility distribution on the Cartesian product {d_i, ¬d_i} x {e_j, ¬e_j} by the following table:

TABLE II. POSSIBILITY DISTRIBUTION ON THE SET OF ELEMENTS E

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

V. Illustrate 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:

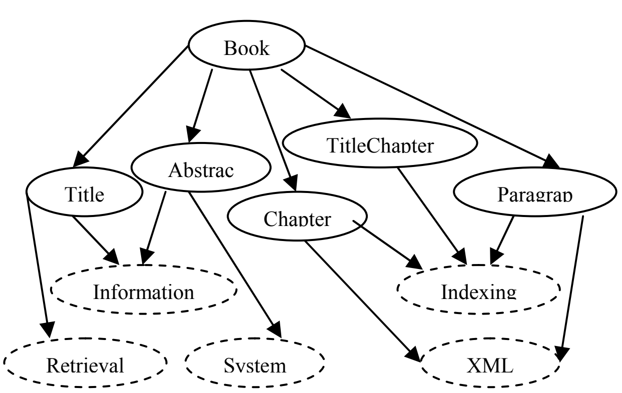

Figure 2 . Possibilistic network of the XML document ‘Book'

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 table containing the values of the arcs element node-term node of the possibilistic network of the document "Book" is given in Table III. We recall that a term is connected to the node that includes it as well as to all the ancestors of this node.

TABLE III. POSSIBILITY DISTRIBUTION Π (t_i/e_j)

Π (t_i/e_j)	t_1	t_2	t_3	t_4	t_5
e_1 $\neg e_1$	1 0	1 0	0 1	0 1	0 1
e_2 $\neg e_2$	0.5 0.5	1 0	1 0	0.25 0.88	0 1
e_3 $\neg e_3$	0 1	0 1	0 1	0.70 0.10	0.5 0.2
e_4 $\neg e_4$	0 1	0 1	0 1	1 0	0 1
e_5 $\neg e_5$	0 1	0 1	0 1	0.88 0.05	1 0

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

TABLE IV. POSSIBILITY DISTRIBUTION 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
$\neg e_1$	1	1
e_2	1	0.1
$\neg e_2$	1	1
e_3	1	1
$\neg e_3$	1	1
e_4	0.6	0.01
$\neg e_4$	1	1
e_5	0.6	1
$\neg e_5$	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 equation of propagation used is the (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, $\forall$i.

Given the query Q, the propagation process (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, e_2), (e_1, ¬e_2), (¬e_1, e_2), (¬e_1, ¬e_2)}. We calculate then:

For d_i = book:

$a_1 = \Pi(t_1/e_1) . \Pi (t_2/e_1). \Pi (t_2/e_2) . \Pi (e_1/book) .$
$\Pi (e_2/ book) = 1 * 1*1 * 1 * 1= 1$
$a_2 = \Pi(t_1/e_1) . \Pi (t_2/e_1) . \Pi (t_2/\neg e_2) . \Pi (e_1/book) .$
$\Pi (\neg e_2/book) = 1* 1*0 * 1 * 1= 0$
$a_3 = \Pi (t_1/\neg e_1) . \Pi (t_2/\neg e_1) . \Pi (t_2/e_2) . \Pi (\neg e_1/book) .$
$\Pi (e_2/book) = 0 * 0*1 * 1 * 1 = 0$
$a_4 = \Pi(t_1/\neg e_1). \Pi(t_2/\neg e_1). \Pi (t_2/\neg e_2). \Pi (\neg e_1/book) .$
$\Pi (\neg e_2/ book) = 0 * 0 * 0 * 1 * 1 = 0$

According to (3):
$\Pi(Q/book) = max (a_1, a_2, a_3, a_4) = 1 = a_1$

For ¬ d_i = ¬ book:

$a_5 = \Pi (t_1/e_1). \Pi (t_2/e_1). \Pi(t_2/e_2). \Pi(e_1/\neg book) .$
$\Pi(e_2/\neg book) = 1 * 1 * 1 * 0.02 * 0.1 = 0.002$
$a_6 = \Pi (t_1/e_1). \Pi (t_2/e_1). \Pi (t_2/\neg e_2). \Pi(e_1/\neg book) .$
$\Pi(\neg e_2/\neg book) = 1 * 1* 0 * 0.02 * 0.1 = 0$
$a_7 = \Pi (t_1/\neg e_1) . \Pi (t_2/\neg e_1). \Pi (t_2/e_2) . \Pi(\neg e_1/\neg book) .$
$\Pi(e_2/\neg book) = 0 * 0 * 1 * 1 * 0.1 = 0$
$a_8 = \Pi(t_1/\neg e_1).\Pi(t_2/\neg e_1).\Pi(t_2/\neg e_2) . \Pi(\neg e_1/\neg book) .$
$\Pi(\neg e_2/\neg book) = 0 * 0* 0 * 1 * 1 = 0$

According to (3):
$\Pi(Q/\neg book) = max (a_5, a_6, a_7, a_8) = 0.002 = a_5$

The necessity N(Q/book) =1- Π(Q/¬ book) = 1- 0.002 = 0.998
The necessity N(Q/¬ book) =1- $\prod$(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.

VI. EXPERIMENTS AND RESULTS

A. Goals

All studies performed to assess aggregated search were based on user studies [14].

This user study is designed with two major goals:

- Evaluate the aptitude of an aggregate of XML elements to answer user queries.
- Identify some of the advantages of the aggregated search in XML documents.

B. Results

To evaluate our model, a prototype was developed. Our experiments are conducted on a sample about 2000 XML documents of the INEX'2005 collection, a set of 20 queries from the same collection. Every query is assessed by exactly 3 users.

The following histogram gives the judgments of users by query regarding the aggregate relevance:

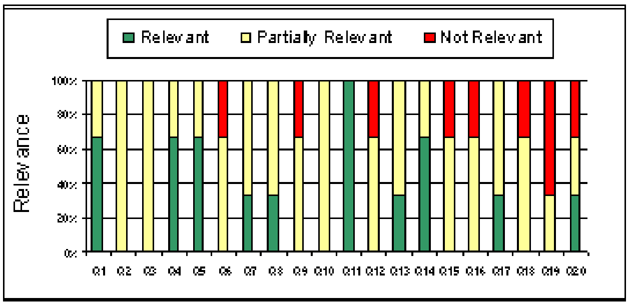

Figure 3 . Distribution of aggregation relevance results

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.

VII. CONCLUSION

This paper presents a new approach for aggregated search based on possibilitic networks. This model provides a formal setting to aggregate non-redundant elements into the same unit. It also directs the user more quickly toward the relevant elements of XML document.

The user study is constructed around 2 main goals. First, we analyze the distribution of relevant results across different elements of XML document. Second, we identify some of the advantages of aggregated search. The user study was used to collect supporting data for these goals. The analysis of the distribution of relevant results provides interesting information. We notice that relevant information is sparse across many elements of XML document.

Our analysis focuses on specific advantages of aggregated search. It is shown that aggregated search is useful to identify different interpretations of a query. It helps find different aspects of the same information need.

Thus, it seems very important to identify other evaluation criteria to identify all benefits of aggregated search in XML documents.

REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, "Diversifying search results", ACM Int. Conference on WSDM, 2009.

[2] N. Ben Amor, "Qualitative possibilistic graphical models: from independence to propagation algorithms", Thèse pour l'obtention du titre de Docteur en Gestion, université de Tunis, 2002.

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

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

[5] A. Brini, M. Boughanem and D. Dubois, "A model for information retrieval based on possibilistic networks", SPIRE'05, Buenos Aires, LNCS, Springer Verlag, pp. 271-282, 2005.

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

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

[8] D. M. Dunlavy, D. P. O'Leary, J. M. Conroy and J. D. Schlesinger, "QCS: A system for querying, clustering and summarizing documents", Information Processing and Management, pp. 1588-1605, 2007.

[9] N. Fuhr, M. Lalmas, S. Malik and Z. Szlavik, "Advances in XML information retrieval: INEX 2004", Dagstuhl Castle, Germany, 2004.

[10] Y. Huang, Z. Liu and Y. Chen, "Query biased snippet generation in XML search", ACM SIGMOD, pp. 315-326, 2008.

[11] B. J. Jansen and A. Spink, "An Analysis of document viewing pattern of web search engine user", Web Mining: Applications and Techniques, pp. 339-354, 2005.

[12] B. J. Jansen and A. Spink, "How are we searching the world wide web?: a comparison of nine search engine transaction logs", Information Processing and Management, pp. 248-263, 2006.

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

[14] A. Kopliku, F. Damak, K. Pinel-Sauvagnat and M. Boughanem, "A user study to evaluate aggregated search", In IEEE/WIC/ACM International Conference on Web Intelligence, in press.

[15] M. Lalmas, "Dempster-Shafer's theory of evidence applied to structured documents: modelling uncertainty", In Proceedings of the 20th Annual International ACM SIGIR, pp.110–118, Philadelphia, PA, USA. ACM, 1997.

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

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

[18] B. Piwowarski, G.E. Faure and P. Gallinari, “Bayesian networks and INEX”, In INEX 2002 Workshop Proceedings, pp. 149-153, Germany, 2002.

[19] N. Polyzotis and M. N. Garofalakis, “XCluster synopses for structured XML content”, ICDE, pp. 63, 2006.

[20] D. Radev, J. Otterbacher, A. Winkel and S. Blair-Goldensohn, “NewsInEssence: summarizing online news topics”, Communications of the Association of Computing Machinery, pp. 95-98, 2005.

[21] T. Rölleke, M. Lalmas, G. Kazai, I. Ruthven and S. Quicker, “The accessibility dimension for structured document retrieval”, BCS-IRSG European Conference on Information Retrieval (ECIR), Glasgow, Mars 2002.

[22] K. Sauvagnat, “Modèle flexible pour la recherche d’information dans des corpus de documents semi-structurés”, Thèse de Doctorat de l’Université Paul Sabatier, Juillet 2005.

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

[24] A. Spink, B. J. Jansen, D. Wolfram and T. Saracevic, “From e-sex to e-commerce: web search changes”, IEEE Computer Science, vol. 35, pp. 107-109, 2002.

[25] S. Sushmita and M. Lalmas, “Using digest pages to increase user result space: preliminary designs”, Special Interest Group on Information Retrieval 2008 Workshop on Aggregated Search, 2008.

[26] L. A. Zadeh, “Fuzzy sets as a basis for a theory of possibility”, In Fuzzy Sets and Systems, 1:3-28, 1978.

Fatma Zohra Bessai-Mechmache Algiers, Algeria. She obtained her Engineer Degree in Computer Science from Houari Boumediene University (USTHB) in Algeria and her Magister from the Research Centre of Advanced Technologies in Algeria.

She has been a team member of the scientific and research staff of the Research Centre in Scientific and Technical Information of Algeria (CERIST) and from 2007 she is the head of the Databases Team at CERIST. Her research interests include database security and information retrieval. She is particularly interested in XML information retrieval, Aggregated Search and Mobile information retrieval.

Zaia Alimazighi Algiers, Algeria. She obtained Doctorate in Computer Science in 1976 at Paris VI University and PHD in Information System at USTHB (Algiers’s University) in 1999.

After 1976 and during more than ten years, she has been project leader in several industrial projects, in public companies in Algeria. She is a researcher at the USTHB since 1988 to nowadays. Today she is a full prof. at the computer science Department of the Faculty of Electronic & Computer Science of USTHB and Dean of this faculty. She is team manager in Computer Engineering Laboratory at USTHB. The current research interests include Information Systems, Collaborative Works, Data Warehouses, Service Web development and Databases Modeling.