International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:8, 2008

Using Dempster-Shafer Theory in XML Information Retrieval

F. Raja, M. Rahgozar, and F. Oroumchian

***Abstract*—**XML is a markup language which is becoming the standard format for information representation and data exchange. A major purpose of XML is the explicit representation of the logical structure of a document. Much research has been performed to exploit logical structure of documents in information retrieval in order to precisely extract user information need from large collections of XML documents. In this paper, we describe an XML information retrieval weighting scheme that tries to find the most relevant elements in XML documents in response to a user query. We present this weighting model for information retrieval systems that utilize plausible inferences to infer the relevance of elements in XML documents. We also add to this model the Dempster-Shafer theory of evidence to express the uncertainty in plausible inferences and Dempster-Shafer rule of combination to combine evidences derived from different inferences.

***Keywords*—**Dempster-Shafer theory, plausible inferences, XML information retrieval.

I. Introduction

THE eXtensible Markup Language (XML) is becoming a standard document format on the World Wide Web (WWW). The widespread use of XML and the continuous growth in XML data sources on the web has brought up the need for effectively retrieving desired XML data from large collections of XML documents. Taking into account the logical structure of documents, represented by XML markups, allows the retrieval process to focus on those parts of the documents that are most relevant to an information need, especially in long documents and documents that cover a variety of subjects.

In a typical information retrieval (IR) system, a document is the only information unit that is indexed and retrieved by the system, and is presented to the user as a result of his/her query. In contrast, in XML Information Retrieval (XML IR) systems each document is considered as a text document with additional structural markups. XML IR systems exploit these structural hints to retrieve relevant elements instead of a whole document in response to the user query.

F. Raja, M. Rahgozar are with the Database Research Group, Control and Intelligent Processing Center of Excellence, faculty of ECE, School of Engineering, University of Tehran, Tehran, Iran (e-mail: f.raja@ece.ut.ac.ir, rahgozar@ut.ac.ir).

F. Oroumchian is with the College of IT, University of Wollongong in Dubai (e-mail: FarhadO@uowdubai.ac.ae).

Therefore, XML IR is a more focused approach than traditional IR which can reduce the users' effort to locate relevant content by directing them to the most relevant parts of the documents.

On the other hand, due to the complex nature of information, determining the semantic content of a document is a highly uncertain task. So applying theories of plausible inferences such as Human Plausible Reasoning (HPR) [7] in IR can make an IR system capable of capturing and handling relationships and uncertainty hidden in documents and inherent in IR process. Such theories of plausible inferences often introduce some uncertainty parameters to represent the uncertainty value of the relations in the documents.

The notions of evidence and uncertainty are not specific to IR, and frameworks have been developed to formally express them. The one adopted in this model is the Dempster-Shqfer theory of evidence (D-S) [6]. Investigations are necessary to either prove or refuse this theory; however some believes that the D-S theory of evidence is both promising and sufficient for the modeling of uncertainty inherent to the structured documents. The reasons are: (i) expressive IR models based on the D-S framework have been proposed (see [3]); (ii) the theory is particularly appropriate in capturing the uncertainty associated with composite objects because it provides an aggregation operator, the Dempster's combination rule that allows the combination of different evidences resulted from different inferences [4].

The rest of this paper is organized as follows: Section 2 presents Dempster-Shafer's (D-S) Theory of Evidence. The theory is then used to construct a weighting model that takes into account the uncertainty that arises in IR by means of plausible inferences (Section 3). Finally the paper finishes the summary of the work and some suggestions for future work (Section 4).

II. Dempster-Shafer's Theory of Evidence

Dempster-Shafer's (D-S) Theory of Evidence is a theory of uncertainty [8] that was first introduced by statistician Arthur Dempster [2] and extended by Glenn Shafer [6]. Its main difference to probability theory is that it allows the explicit representation of ignorance and the combination of evidence. It is the latter property that makes the use of the D-S theory particularly attractive in modeling the IR process [3]. The combination of evidence is expressed by the *Dempster's combination rule*, which allows the expression of aggregation necessary in a model for structured document representation

and retrieval. Examples of applications of this theory in IR can be found in [1, 3, 4, 5]. In the context of this problem, uncertainty refers to the following three cases: First, the existence of multiple evidences for the relevance of a document to a query term. Second the amount of unspecified evidences for the relevance of that document to the same query term. Third, the misleading evidences that incorrectly identify the document as relevant to that query term. In our problem the evidences are compatible with each other and have no conflict.

In this section we describe the main concepts of D-S theory, based on the description given in [6].

A. Frame of Discernment

The D-S framework is based on the view whereby propositions are represented as subsets of a given set. Suppose that we are concerned with the value of some quantity *u*, and the set of its possible values is *U*. The set *U* is called a *frame of discernment*. An example of a proposition is "the value of *u* is in *A*" for some $A \subseteq U$. Thus, the propositions of interest are in a one-to-one correspondence with the subsets of *U*. The proposition $A = \{a\}$ for $a \in U$ constitutes a basic proposition "the value of *u* is *a*". In our problem the set of all the possible answers to the query, the elements of documents, is our frame of discernments.

B. Basic Probability Assignment

Beliefs can be assigned to propositions to express their uncertainty. The beliefs are usually computed based on a density function $m : \wp(U) \rightarrow [0,1]$ called a *basic probability assignment* (*bpa*):

$$m(\emptyset) = 0 \quad \text{and} \quad \sum_{A \subseteq U} m(A) = 1 \qquad (1)$$

$m(A)$ represents the belief exactly committed to *A*, that is the exact evidence that the value of *u* is in *A*. If there is positive evidence for the value of *u* being in *A* then $m(A) > 0$, and *A* is called a *focal element*. The proposition *A* is said to be *discerned*. No belief can ever be assigned to the false proposition (represented as ∅). The sum of the non-null *bpa*s must equate 1. The focal elements and the associated *bpa* define a *body of evidence*.

C. Dempster's Combination Rule

D-S theory has an operation, *Dempster's rule of combination*, for the pooling of evidence from a variety of sources. This rule aggregates two *independent* bodies of evidence defined within the same frame of discernment into one body of evidence. Let $m1$ and $m2$ be the *bpa*s associated to two independent bodies of evidence defined in a frame of discernment *U*. The new body of evidence is defined by a *bpa* *m* on the same frame *U*:

$$m(A) = m1 \otimes m2(A) = \frac{\sum_{B \cap C = A} m1(B) m2(C)}{\sum_{B \cap C = \emptyset} m1(B) m2(C)} \qquad (2)$$

Dempster's combination rule, then, computes a measure of agreement between two bodies of evidence concerning various propositions discerned from a common frame of discernment. The rule focuses only on those propositions that both bodies of evidence support. The new *bpa* takes into account the *bpa* associated to the propositions in both bodies that yield the propositions of the combined body. The denominator of the above equation is a normalization factor that ensures that *m* is a *bpa*.

III. A New Weighting Model for XML Documents

In this weighting schema we assume that the IR system uses plausible inferences to infer the relevance of elements to query terms. Then since a term in an element could be reached through several different inferences, therefore a method for combining the confidence values attributed from these different evidences should be taken. For this purpose, D-S rule of combination, as described in [1], will be used. It combines the different confidence values estimated by different inferences for each term.

In IR systems that apply plausible inferences every inference returns a confidence value for each element inferred by each term of the query. These confidence values are modeled by *bpa*.

$$m(\emptyset) = 0 \quad \text{and} \quad \sum_{A \subseteq U} m(A) = 1 \qquad (3)$$

m(A) represents the exact evidence that the element is relevant to a query term. If there is positive evidence for relevance of an element to a query term, then m(A) > 0, and A is a focal element. The focal elements and the associated *bpa* define a body of evidence. In this problem, we assume that focal elements are singleton sets. If m(A) = 0 then there is no confidence about relevance of an element to any particular query term. Each body of evidence is composed of the confidence on relevance of an element to each query term as estimated by inferences of plausible reasoning. Then

$$m(\emptyset) = 0 \quad \text{and} \quad m(\{element_j\}) + m(T) = 1 \qquad (4)$$

m(T) is referred to evidence that can not be assigned yet (uncommitted belief as described in [3]). The m(T) represents the uncertainty (overall ignorance, lack of knowledge) associated to the entire set of elements about being relevant to a query term.

Now to combine the evidences of relevance of an element to a query term first we define Ci, Wi (weight of inference i) and mi(Ej) (mass of evidences of relevance for element j from inference i):

Ci = confidence on relevance of each leaf element for each term of the query returned by inference i.

Wi = number of relevant elements retrieved by inference i / total elements retrieved by inference i.

Then:

$$W'_i = \frac{W_i}{\sum_{\#ofInferences} W_i} \quad (5)$$

$$m_i(E_j) = W'_i * C_i \quad (6)$$

$$m(E_j) = m_1(E_j) \otimes m_2(E_j) \otimes \ldots \otimes m_{\#ofInferences}(E_j) \quad (7)$$

Because the focal elements are singleton, the combination function becomes simpler than Dempster's rule of combination and only the evidences with mi(Ej)>0 combine with each other.

$$m(E_j) = m_1(E_j) \otimes m_2(E_j) = (m_1(E_j) * m_2(E_j) + m_1(E_j) * m_2(T) + m_2(E_j) * m_1(T)) / K \quad (8)$$

$$m(T) = m_1(T) * m_2(T) / K \quad (9)$$

Where K is a normalization factor to support that m is *bpa*.

$$K = m_1(E_j) * m_2(E_j) + m_1(E_j) * m_2(T) + m_2(E_j) * m_1(T) + m_1(T) * m_2(T) \quad (10)$$

Where:

$$m(E_j) + m(T) = 1 \quad (11)$$

We combined the evidences of relevance from different inferences for each query term. In the next level, the evidences of relevance from different query terms must be combined to compute the final evidence value for the element. For this purpose we define the weight of each term in each element. The following formula assigns a *bpa* to each term of the query in each element. If we have:

Frame of discernment = {all elements in the collection} =T

Now each term for each element is a source of evidence. In this step we should define the weight of each query term in each element. We propose the following weighting model for this purpose:

$$\begin{cases} m(A) = tf(e,t) * \log_N^{\frac{N}{n(t)}} * cv(e,t) & A = \{e\}, t \in e \\ m(A) = 1 - \sum_{\forall e} m(e) & A = T \end{cases} \quad (12)$$

Where:

$$tf(e,t) = \frac{occ(e,t)}{T(e)} \quad (13)$$

occ(e,t) = number of occurrences of term *t* in element *e*(term frequency of term *t* in element *e*).

And:

$$T(e) = \sum_{\forall e, \forall t} occ(e,t) \quad (14)$$

N= number of total elements existing in the collection.

n(t)= number of elements containing term *t*

cv(e,t) = confidence value of relevance of element *e* to query term *t* as computed in (7).

As can be seen, the two first factors in computing m(A) ,where A={e}, are term frequency (*tf*) and inverse document frequency (*idf*), respectively. "*tf*" is an element-specific statistic; it varies from one element to another, attempting to measure the importance of the term within a given element. By contrast, "*idf*" is a global statistic which characterizes a given term within an entire collection of elements. It is a measure of how widely the term is distributed over the given collection, and hence the fewer the elements containing the given term, the larger the "*idf*". So, a term that occurs in every element in the collection is not likely to be useful for distinguishing relevant from non-relevant elements.

Now that we have the weight of each query term for each element, first we should show that each weight is a *bpa* and then combine the weights that all query terms assign to an element to reach the final value of relevance of that element to the user query.

Now we show that each weight is a *bpa* and satisfies the D-S conditions:

$$\left.\begin{array}{l} 0 \le occ(e,t) \le T(e) \Rightarrow 0 \le f(e,t) \le 1 \\ 1 \le n(t) \le N \Leftrightarrow 1 \le \frac{N}{n(t)} \le N \Leftrightarrow 0 \le \log_N \frac{N}{n(t)} \le 1 \\ cv(e,t) = bpa \Rightarrow 0 \le cv(e.t) \le 1 \end{array}\right\}$$

$$\Rightarrow 0 \le m(A) \le 1 \Rightarrow m(A)$$

$$\sum_{A \subseteq U} m(A) = 1$$

To find the final value of relevance of an element to a query, we combine the weights that all query terms assign to an element as follows:

$$m(e) = m_1(e) \otimes m_2(e) = (m_1(e) * m_2(e) + m_1(e) * m_2(T) + m_2(e) * m_1(T)) / K \quad (15)$$

$$m(T) = m_1(T) * m_2(T) / K \quad (16)$$

$$K = \sum_{\forall e} (m_1(e) * m_2(e) + m_1(e) * m_2(T) + m_2(e) * m_1(T) + m_1(T) * m_2(T)) \quad (17)$$

Where:
mi(e) = the weight that term i assigns to element e.
mi(T) = the weight that term i assigns to T.

Finally since we start this process from leaf elements in the document tree, we should combine the relevance values of child nodes of a non-leaf element in order to find its relevancy to the user query. We can use the above D-S theory for in this step also. Now that we have the relevancy of all elements in the collection we can rank them based on the total relevance values and represent to the user.

IV. Summary and Future Works

In this paper we presented a new weighting model for XML IR systems which apply plausible inferences to infer about the relevance of an element in an XML document to a user information need. In this model we use Dempster-Shafer rule of combination in three steps; first, we combine the different confidence values attributed from different inferences for each query term in each leaf element; then, we use effective parameters in weighting models in IR systems (*tf* and *idf*) to find a total weight for each query term. The most innovative part of this weighting model is the way we adopt our term weight to *bpa*s in D-S theory. By defining term weights as *bpa*s we can combine different weights for different terms in the query using D-S rule of combination; after that, we use D-S rule of combination to combine relevance values of child nodes of non-leaf elements to find the relevancy of all elements of the XML document.

This weighing scheme seems to be an effective weighting scheme since it considers the confidence values of plausible inferences as well as the most important parameters of weighting models in IR. On the other hand, it benefits from the advantages of D-S rule of combination in structured documents.

In future we plan to implement these ideas in an XML IR system and compare it with other weighting schemas proposed in this field.

References

[1] F. Orumchian, B. Nadjar Araabi, and E. Ashoori, "Using Plausible Inferences and Dempster-Shafer Theory of Evidence for Adaptive Information Filtering", 4th International Conference on Recent Advances in Soft Computing, Nottingham, United Kingdom - 2002.

[2] A. Dempster, "Upper and lower probabilities induced by a multivalued mapping," Ann. Math. Stat., vol. 38, no. 2, pp. 325–339, 1967.

[3] M. Lalmas, and M. Ekaterini, "A Dempster-Shafer indexing for focussed retrieval of a hierarchically structured document space: Implememtation and experiments on a web museum collection," 6th RIAO Conference, Content-Based Multimedia Information Access, Paris, France, April, 2000.

[4] I. Ruthven, and M. Lalmas, "Using Dempster-Shafer's Theory of Evidence to combine aspects of information use," Journal of Intelligent Information Systems, 2001

[5] I. Ruthven, and M. Lalmas, "Experimenting on Dempster-Shafer's theory of evidence in information retrieval," Technical report, University of Glasgow, April 1998.

[6] A G.A. Shafer, "Mathrematical Theory of Evidence", Princeton University Press, 1976.

[7] A. Collins and R. Michalski. "The logic of plausible Reasoning A core theory", cognitive science, vol. 13, pp.1-49, 1989.

[8] A. Saffioti, "An AI view of the treatment of uncertainty",The Knowledge Engineering Review, 2(2), 1987, 75-97.