Evidential-Link-based Approach for Re-ranking XML Retrieval Results

M'hamed Mataoui[1,2], Mohamed Mezghiche[2], Faouzi Sebbak[3] and Farid Benhammadi[3]
[1]*IS&DB laboratory, Ecole Militaire Polytechnique, Bordj el Bahri, Algiers, Algeria*
[2]*LIMOSE laboratory, M'hamed Bougara University of Boumerdes, Boumerdes, Algeria*
[3]*AI laboratory, Ecole Militaire Polytechnique, Bordj el Bahri, Algiers, Algeria*

Keywords: Topic-sensitive, Query Dependent, Re-ranking Approach, XML Information Retrieval, XML links, Link Analysis Algorithms, INEX.

Abstract: In this paper, we propose a new evidential link-based approach for re-ranking XML retrieval results. The approach, based on Dempster-Shafer theory of evidence, combines, for each retrieved XML element, content relevance evidence, and computed link evidence (score and rank). The use of the Dempster–Shafer theory is motivated by the need to improve retrieval accuracy by incorporating the uncertain nature of both bodies of evidence (content and link relevance). The link score is computed according to a new link analysis algorithm based on weighted links, where relevance is propagated through the two types of links, i.e., hierarchical and navigational. The propagation, i.e. the amount of relevance score received by each retrieved XML element, depends on link weight which is defined according to two parameters: link type and link length. To evaluate our proposal we carried out a set of experiments based on INEX data collection.

1 INTRODUCTION

New challenges in information retrieval (IR) field have appeared by the growing quantity of available structured information resources, principally collections of XML documents. Therefore, the logical (hierarchical) structure of XML documents, representing a new source of evidence, is exploited to retrieve XML elements at different levels of granularity. Instead of classical information retrieval approaches that focus on seeking unstructured content, XML information retrieval combines both textual and structural information to perform different IR tasks. A number of approaches taking advantage of the two types of information (textual and structural) have been proposed and are essentially based on traditional information retrieval models adapted to process the content part of the XML documents context (Fuhr and Großjohann, 2001; Guo et al., 2003; Kimelfeld et al., 2007).

Despite the popularity of links in the web (Guo et al., 2003; Kamps and Koolen, 2008; Kimelfeld et al., 2007; Pehcevski et al., 2008; Zhang and Kamps, 2008) and the conceptual proximity between HTML and XML links, only few of IR approaches have exploited links connecting XML documents in XML IR context. Hyperlinks have been used by several well-known algorithms, including PageRank (Brin and Page, 1998), HITS (Kleinberg, 1999) and SALSA (Lempel and Moran, 2001), to evaluate page relevance with respect to user query. XML IR approaches (Kamps and Koolen, 2008; Kimelfeld et al., 2007; Pehcevski et al., 2008; Zhang and Kamps, 2008) exploiting XML links were adapted from these well-known web-based algorithms by assigning link scores to documents instead of XML elements by the consideration of hyperlinks at document granularity, i.e.. This could be because of the links form in the used collection, for example, in one of the main XML test collections, namely, INEX Wikipedia collection (Denoyer and Gallinari, 2007), links point to the root of XML documents instead of internal elements.

Based on the well-known mathematical theory of Dempster–Shafer theory (also known as belief function theory),some approaches have been proposed in the literature. Lalmas and Ruthven (Lalmas and Ruthven, 1998) used the DS theory of evidence to combine aspects of information use. The proposed model combines evidence from user's relevance with algorithms describing how words are used withing documents. They also present some experimenting on this theory in information retrieval. Schocken and Hummel (Schocken and Hummel, 1993) used DS theory to combine taxonomies of keywords. In their

Mataoui M., Mezghiche M., Sebbak F. and Benhammadi F..
Evidential-Link-based Approach for Re-ranking XML Retrieval Results.
DOI: 10.5220/0005003900640071
In *Proceedings of 3rd International Conference on Data Management Technologies and Applications* (DATA-2014), pages 64-71
ISBN: 978-989-758-035-2

approach different confidence levels are assigned for each defined keyword set. Then, using DS theory, they combine these assignments to find the new mass distribution over these sets. The use of this theory is mainly motivated by the incorporation of the uncertain nature of information retrieval.

In this paper, we propose an evidential-link-based approach for re-ranking XML retrieval results. This approach is based on a combination of textural and structural information. To evaluate our proposal we have conducted a series of experiments on the INEX collection devoted to XML IR evaluation.

This paper is organized as follows: In Section 2, related work is presented. Section 3 describes our approach aiming at exploiting the different types of links and the weight of links between elements in XML IR. Section 4 presents the Dempster–Shafer (DS) theory as well as its application in IR field. Section 5 shows the experimental results, we focused on the comparative experiments and discussed our findings after evaluation. Finally, in Section 5, we conclude with some prospects.

2 RELATED WORK

Few researches have been conducted in the XML information retrieval context to exploit link evidence. These researches can be classified into three classes: (a) approaches analyzing the structure and nature of links in XML documents collections (Kamps and Koolen, 2008; Zhang and Kamps, 2008); (b) approaches based on the link detection strategies, often called "Link-The-Wiki" task in INEX initiative (Dopichaj et al., 2009; Geva et al., 2009; Itakura et al., 2011; Jenkinson et al., 2009; Fachry et al., 2008; Zhang and Kamps, 2008); and (c) approaches exploiting links to re-rank the initially list of XML elements returned by retrieval systems (Kamps and Koolen, 2008; Kimelfeld et al., 2007; Pehcevski et al., 2008; Zhang and Kamps, 2008). In this section, we focused on the last class.

Guo et al. proposed XRANK (Guo et al., 2003), one of the first works which exploits XML links as a source of evidence in the computation of retrieved XML elements relevance scores. The computation of the link score is based on three types of links between XML nodes. XRANK suffers from several limits. First, the proposed link score computation formula is used exclusively in entire collection context which does not improve the retrieval accuracy. The second limit is that XRANK cannot be exploited in the topical context. Finally, several of XML IR tasks do not allow overlapping, which make no sense to the proposed formula for these XML IR tasks. The XRANK approach was evaluated upon two datasets: XMARK and DBLP. The only performed experiment upon the XRANK retrieval system was related to the performance factor, i.e., execution time, and not to the retrieval accuracy.

After the advent, in 2002, of the INEX initiative for the Evaluation of XML retrieval (Gövert and Kazai, 2002), more works have been proposed to exploit the XML links. Kimelfeld et al. (Kimelfeld et al., 2007) applied HITS algorithm (Kleinberg, 1999) upon the top-N retrieved XML documents to filter returned results. Obtained evaluation results have not been convincing and authors proposed, as prospects, to use Pagerank instead of HITS. In our previous work (Mataoui et al., 2010), we showed also that using HITS on INEX 2007 collection does not improve retrieval effectiveness but rather contrary. Kamps J. and Koolen M. (Kamps and Koolen, 2008), Fachry et al. (Fachry et al., 2008) exploited two levels: "global indegree" and "local indegree" of the XML links to re-rank the retrieval results. This approach is specific to document level granularity (document-to-document link type) and can induce in error because, in general, the number of incoming links does not give a precise vision of the XML document relevance, but its link quality. For instance, a document pointed to by only one link from a highly relevant document can be more relevant compared to another document pointed to by many incoming links from irrelevant documents. Philippe Mulhem and Delphine Verbyst (Verbyst and Mulhem, 2009) describe a method to incorporate link score in the computation of the final score of Doxels (XML elements). Their approach is based on both exhaustivity and specificity scores between linked doxels. The proposed formula is applied in a global context. Authors showed, by experiments on the INEX XML collection, that "element-element" link type can improve retrieval accuracy.

All these, earlier mentioned, link based approaches, excepting XRANK and Doxels approaches, do not propose solutions based on the "element-element" link type.

Contrary to the previous works the approach we present in this paper attempts to exploit "element-element" links (path), composed either by internal (hierarchical) and/or external (navigational) links. Since most of XML collections contains "element-document" link type, we propose a solution that allows to propagate "element-document" link to the elements of the target document. In addition, the proposed approach in this paper uses th DS theory to combine initial results scores extracted from INEX

data collection with the computed link scores by the new "topic-sensitive" XML IR approach.

3 WEIGHTED LINKS BASED APPROACH

We propose, in this paper, an evidential "topic-sensitive" approach that combines both initial content relevance score and link evidence score to compute a new relevance score for each retrieved XML element. The new computed relevance score is used to re-rank the initial retrieved list of XML elements. We focused in this paper on the manner XML links, both navigational and hierarchical links could be used to compute link evidence score of retrieved XML elements.

To introduce the way the link score is computed we define a hyperlinked collection of XML elements returned as retrieval results for a given topic Q as a directed graph $\Omega = (Q, E, NLTG, HLTG)$; where Q represents the topic (query) for which retrieved XML elements are returned as response; E represents the nodes of the graph, i.e., the set of retrieved XML elements in response to Q; *NLTG* represents the navigational (external) links and *HLTG* the hierarchical (internal) links between XML elements belonging to E. Navigational links are supposed as unidirectional links and hierarchical as bidirectional links. We explore principally the popularity propagation model exploited in web link analysis algorithms.

We assume that each retrieved XML element has a given relevance score that can be propagated through links. In our approach we interpret the amount of relevance score propagated between two XML elements, *E1* and *E2*, as the probability to explore this path by a user. The propagated amount of relevance is inversely proportional to the path weight. Therefore, the more the path weight between two XML nodes is great, the more the probability to explore this path by a user is less. In our context a path consist of 0 or 1 navigational link and a set of hierarchical links. By considering that it is easier for a user to navigate through navigational (click on the link) than hierarchical links, we assume that the probability that a user traverses a path containing an navigational link is higher than that of a user traverses a path which contains only hierarchical links. Consequently, the propagated relevance depends on the existence of navigational link and the number of links. We call this concept: weighting of the links, where we define a parameter λ that reflects the weight of navigational links *(NLW)* compared to hierarchical links *(HLW)*. We propose the following formula:

$$NLW = \lambda * HLW / \lambda \in]0,1] \quad (1)$$

Increasing of λ value implies increasing of hierarchical links weight.

The algorithm of computation of the path weight is shown in the algorithm 1. As aforementioned, we consider in our approach the two types of links: navigational links *(NL)* and hierarchical *(HL)*. Navigational links connect generally between XML nodes belonging to different XML documents and hierarchical links represent the structure of these documents. As we have mentioned, our approach is applied in "topic-sensitive" context, which means that we exploit a sub-graph of the global link graph. This sub-graph can be obtained by incorporating two entities, which are: retrieval results and global link graph. To obtain the "topic-sensitive" link graph we extract the two link-type graphs as shown in figure 1 and 2.

Algorithm 1: Path Weight "$PW(N_i, N_j)$" Computation Algorithm.

if $\exists EP / (N_i \rightarrow EP)$ *is a navigational link* **and** $(N_i \rightarrow N_j) \equiv (N_i \rightarrow EP) \cup (EP \rightarrow N_j)$ **then**
  $PW(N_i, N_j) \leftarrow [NLW + [dist(EP, N_j) * HLW]]$
else
  $PW(N_i, N_j) \leftarrow [dist(EP, N_j) * HLW]$
end if

To illustrate how "Path Weight" information is used to compute link scores of the retrieved XML elements we take the example of figure 1. Let a link graph containing four XML documents: "document1.xml", "document2.xml", "document3.xml" and "document4.xml". These documents contain five retrieved elements for a given query Q: *Node1*, *Node2*, *Node3*, *Node4* and *Node5*. These XML elements are connected by 3 navigational links *NL1*, *NL2* and *NL3*. We notice that *Node3* and *Node4* can be reached from *Node1* by traversing *NL1*. *Node3* and *Node4* can also be reached from *Node2* by traversing *NL2*. *Node5* can be reached from *Node3* by navigating through *NL3*. *Node3* can be reached from *Node4* and *Node4* from *Node3* by navigating through hierarchical structure of "document3.xml".

Figure 2 represents a subgraph of the link structure of figure 1 where only retrieved elements and their links weighted according to algorithm 1 are mentioned.

We now consider the problem of computing link scores of XML elements. As mentioned, the link score is a measure of the XML element importance, and it is computed based on the topic-sensitive link graph, i.e., retrieved XML elements. To compute the amount of propagated relevance score that passes through the link structure connecting two XML nodes N_i to N_j, we propose the formula of equation 2 taking into account the two types of links and the path weight

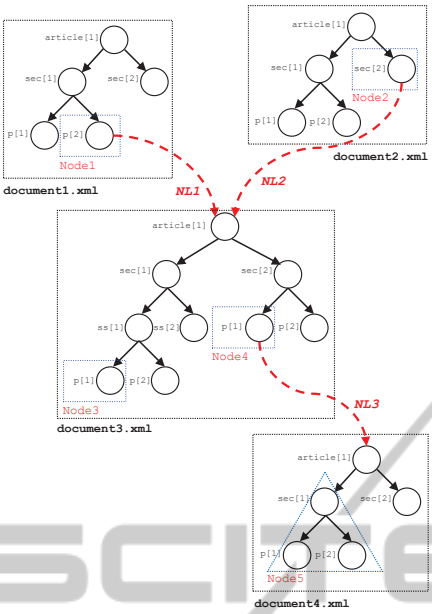

Figure 1: Example of link structure graph (hierarchical and navigational links).

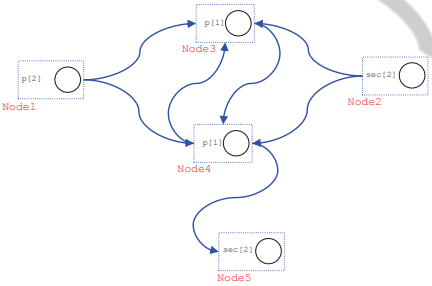

Figure 2: "Topic-sensitive" link graph construction for the example of figure 1.

between these XML elements.

To formalize this propagation process, we consider $RS(N_i)$ as the current relevance score of XML node N_i, and $URS(N_i)$ as the unit of propagated relevance score by N_i through a path with $PW = 1$. $PRS(N_i \rightarrow N_j)$ represents the propagated relevance score by XML node N_i to N_j. $PW(N_i \rightarrow N_j)$ represents the weight of the path between N_i and N_j computed according to algorithm 1.

$$\begin{cases} PRS(N_i \rightarrow N_j) \leftarrow \dfrac{URS(N_i)}{PW(N_i \rightarrow N_j)} \\ \displaystyle\sum_{N_j \in Outlinks(N_i)} PRS(N_i \rightarrow N_j) = RS(N_i) \end{cases} \quad (2)$$

Second part of equation 2 represents the constraint related to the sum of the amount of relevance scores propagated by a given XML node which must not exceed (be equal) to its own relevance score. We define N_j as the set of XML nodes reached from outlinks of XML node N_i. Only active outlinks, i.e., those pointing to retrieved elements are considered. Equation 3 represents the way the unit of propagated relevance score by N_i through a path (with PW=1) is computed.

$$\sum_{N_j \in Outlinks(N_i)} PRS(N_i \rightarrow N_j) = RS(N_i)$$

$$\Rightarrow URS(N_i) = \frac{RS(N_i)}{\sum_{N_j \in Outlinks(N_i)} \frac{1}{PW(N_i \rightarrow N_j)}} \quad (3)$$

The final link score *"LS(XE)"* of an XML element XE is computed following equation 4. *"LS(XE)"* is obtained by combining equations 2 and 3, i.e., by summing propagated relevance scores through different links, as follows:

$$LS(XE) = \frac{(1-\rho)}{|N|} + [\rho * \sum_{N_i \in Inlinks(XE)} PRS(N_i \rightarrow XE)]$$

$$\Rightarrow LS(XE) = \frac{(1-\rho)}{|N|} + [\rho * \sum_{N_i \in Inlinks(XE)} \frac{\frac{RS(N_i)}{\sum_{N_j \in Outlinks(N_i)} \frac{1}{PW(N_i \rightarrow N_j)}})}{PW(N_i \rightarrow XE)}] \quad (4)$$

Where:

- $|N|$ represents the number of retrieved XML elements (nodes in the topic-sensitive link graph);
- ρ parameter represents the damping factor (generally fixed at 0.85).

$\frac{(1-\rho)}{|N|}$ represents the probability of visiting randomly an XML element E in the graph of links. The second fragment of equation 4 represents the probability of reaching E by navigating through both link types from other XML elements. Computation of *LS(XE)* is carried out according to an iterative process until the convergence of link scores. Convergence proof of equation 4 can be found in (Farahat et al., 2006). Equation 4 is conceptually comparable to Pagerank, excepting that: (a) the two types of links (navigational and hierarchical) are taking into account in the computation of link score; (b) it exploits a new parameter, namely, path weight in the relevance propagation process; (c) link scores are computed at XML element granularity instead of document granularity; (d) the approach is applied at "topic-sensitive" context, i.e., query dependent;

4 DEMPSTER–SHAFER AND INFORMATION RETRIEVAL

4.1 Introduction

The Dempster-Shafer (DS) theory (known as belief functions) is a theory of uncertainty that was devel-

Table 1: A simple demonstrative worked example.

Element	S_1 *Initial I.R. source*		S_2 *Link I.R. source*		$\alpha_{S_j}(e_i)$		Combined initial masses	Combined discounting masses
	Initial score	Rank	Link score	Rank	*s1*	*s2*		
e_1	0.7	1	0.6	1	1	1	0.778 (1)	0.778 (1)
e_2	0.15	2	0.02	4	0.75	0.25	**0.004 (4)**	**0.089 (3)**
e_3	0.1	3	0.08	3	0.5	0.5	**0.010 (3)**	**0.049 (4)**
e_4	0.05	4	0.3	2	0.25	0.75	0.022 (2)	0.186 (2)

oped by Dempster (Dempster, 1967) and further extended by Shafer (Shafer, 1976). This theory improves quantifying uncertainty by allowing the explicit representation of ignorance. It has attractive properties providing richer information in combining sources of evidence. The DS theory have been used to model various aspects of the information retrieval process (Schocken and Hummel, 1993; Lalmas and Ruthven, 1998).

4.2 DS Theory Elements

The DS theory is based on the grounds of the following concepts and principles:

(a) **The Frame of Discernment** is a set of mutually exclusive and exhaustive hypotheses about the problem domains. From a frame of discernment (Θ) correspondingly 2^Θ is the power set of (Θ).

(b) **A Basic Belief Assignment** *(bba)* or *mass function* represents the degree of belief and is defined as a mapping $m(\cdot)$ satisfying the following properties: $m(\varnothing) = 0$, $\varnothing$: the empty set
$\sum_{H\in 2^\Theta} m(H) = 1$, H: a subset of Θ
The subsets H of the power set 2^Θ with a positive mass of belief is called *focal* set element of $m(\cdot)$.

(c) **The Dempster's Combination Rule** is the most important tool of the evidence theory. This rule aims to aggregate evidence from multiple independent sources defined within the same frame of discernment.

Let m_1 and m_2 be the mass functions associated with two independent bodies of evidence. H_1 and H_2 represent the focal elements of m_1 and m_2 respectively. The mass function m is formed by combining m_1 and m_2 as $m = m_1 \oplus m_2$. This rule with two sources, $m = m_1 \oplus m_2$ is defined by equation 5.

$$m^{DS}(H) = \frac{m_{12}(H)}{1 - m_{12}(\varnothing)} \tag{5}$$

where

$$m_{12}(H) = \sum_{\substack{H_1,H_2\in 2^\Theta \\ H_1\cap H_2 = H}} m_1(H_1)m_2(H_2)$$

Where $m_{12}(H)$ and $m_{12}(\varnothing)$ represent the conventional conjunctive consensus operator and the conflict of the combination between the two sources respectively. Additionally, from a given *bba m*, the belief and the plausibility functions are used as decision criteria (Dempster, 1967).

4.3 The Discounting of Sources of Evidence

It is possible to discount an unreliable source proportionally to its corresponding reliability factor according to the method proposed by Shafer (Shafer, 1976). Shafer assumes that if we know the reliability/confidence factor α that belong to the interval [0,1], then the discounting of the *bba* $m(\cdot)$ provided by the unreliable source denoted by $m'(\cdot)$ is defined as follow:

$$\begin{cases} m'(A) = \alpha.m(A), & \forall A \in 2^\Theta, A \neq \Theta \\ m'(\Theta) = (1-\alpha) + \alpha.m(\Theta) \end{cases} \tag{6}$$

4.4 Using the DS Theory in IR Field

Within the context of information retrieval and according to the proposed new "topic-sensitive" approach, we define the frame of discernment by: $\Theta = \{e_i, \neg e_i\}$, where e_i is a retrieved element. Let S_1 and S_2 be initial and link information retrieval sources respectively. Then, we define two basic belief assignments for initial and link scores obtained from S_1 and S_2 as follows: $m_{S_i}(\varnothing) = 0$, $\varnothing$: the empty set
$\sum_{H\in 2^\Theta} m_{S_i}(H) = 1$, H : a subset of Θ and $S \in \{S_1, S_2\}$

Initial and link scores can be scaled to fall between 0 and 1 in order to satisfy the mass properties as follows:

$$\begin{cases} m_{S_1}(e_i) = \frac{IS(e_i)}{\sum_{j=1\cdots n}(IS(e_j))} \\ \\ m_{S_2}(e_i) = \frac{LS(e_i)}{\sum_{j=1\cdots n}(LS(e_j))} \end{cases} \tag{7}$$

Where n denotes the number of elements.

For XML elements classification decision making, we adopt the combination of initial and link information retrieval scores. This combination is based on Dempster's rule to obtain a final score mass of the returned XML elements.

Let the initial score masses of the retrieved elements for a given query Q be: $m_{S_1}(e_1)$, $m_{S_1}(e_2)$, $\cdots$, $m_{S_1}(e_n)$ and the computed link score masses be: $m_{S_2}(e_1)$, $m_{S_2}(e_2)$, $\cdots$, $m_{S_2}(e_n)$. Then, the combined score mass using Dempster's rule is defined as:

$$m^{DS}(FS) = m_{s_1}(IS) \oplus m_{s_2}(LS) \tag{8}$$

$$m^{DS}(FS(e_i)) = \frac{m_{12}(e_i)}{1 - m_{12}(\emptyset)} \tag{9}$$

$$\text{where} \quad m_{12}(e_i) = \sum_{\substack{H_1,H_2 \in 2^\Theta \\ H_1 \cap H_2 = e_i}} m_{S_1}(H_1) m_{S_2}(H_2)$$

The preceding combination rule does not take into account the discounting factor of the two sources. To deal with the discount problem, we propose a novel discounting method, which can maximize for a given query, a scoring function that implicitly imposes an ordering on documents, directly defined on the rank performance measures. As a result, our discount approach uses a query-dependent ranking model to discount its score. According to each source, this method computes discounting factor of each element (e_i) on the basis of its rank because the ranking measure plays an important role in almost all activities related to information retrieval.

When a new query is consulted, the individual element rank in respect to the source S_j is obtained which is then used to compute the corresponding element discounting factor. This discounting factor is defined by the following formula:

$$\alpha_{S_j}(e_i) = \frac{1}{r_{S_j}(e_i)} \tag{10}$$

where $r_{S_j}(e_i)$ denotes the rank of the element e_i according to their relevance to the query for the user in respect to the source S_j.

Hence, using the Shafer's discounting of each source of evidence S_j and its corresponding factor $\alpha_{S_j}(e_i)$, we proceed to calculate the reliability of each score mass of the element e_i which is defined as follow:

$$m'_{S_j}(e_i) = \alpha_{S_j}(e_i) \cdot m_{S_j}(e_i) \tag{11}$$

$$m'_{S_j}(\neg e_i) = \alpha_{S_j}(e_i) \cdot m_{S_j}(\neg e_i) \tag{12}$$

$$m'_{S_j}(\Theta) = (1 - \alpha_{S_j}(e_i)) + \alpha_{S_j}(e_i) \cdot m_{S_j}(\Theta) \tag{13}$$

Now for each element e_i, we apply Dempster's rule for combining their discounting initial and link scores. This is defined by the following equation:

$$m^{DS}_{(S_1,S_2)}(e_i) = m'_{S_1}(e_i) \oplus m'_{S_2}(e_i) \tag{14}$$

The final scores $m^{DS}_{(S_1,S_2)}(e_i)$ for $i = 1 \cdots n$ allow the re-rank of the initially returned list of XML based on DS theory that use the two "element-element" link types and fixed discounting rates according to the rank function of the elements. Apparently, a higher final score value is better since more relevant documents are placed in front positions.

To show the utility and the effectiveness of these discounting rates in the combination process, let consider the query Q which is associated with four documents (e_1, e_2, e_3, e_4) as reported in Table 1. As can been see, the combined discounting masses for the element e_1 confirms the relevance of this element because each source has ranked e_1 at the first position. However, the element e_2 has been re-ranked (from the fourth rank to the second one) due to its relevance according to the initial information source (s_1) where its score is greater than the score of the element e_3 which is re-ranked at the fourth position.

5 EXPERIMENTATION

5.1 Experimental Setup

Our experiments were performed using INEX 2007 Wikipedia XML collection (Denoyer and Gallinari, 2007; Gövert and Kazai, 2002; Geva et al., 2010; ref, 2013). This collection contains 659,388 XML documents and characterized by its densely and semantically related hyperlinked structure that differs from the Web link structure.

As abovementioned, our approach exploits the re-ranking principal, upon the initially retrieval results returned by an XML retrieval system, by combining the initial relevance score with the computed link score using Dempster–Shafer theory.

To evaluate our proposals, we exploit retrieval results (for "Focused" task) from the three best XML retrieval systems of INEX 2007, namely, Dalian, Waterloo and MaxPlanck systems. These retrieval results related to the 107 CAS (Content And Structure) topics of INEX 2007 (Geva et al., 2010). The INEX "Focused" task focuses on the most specific XML elements. The metric used in this task is the interpolated Precision at 1% level of recall (iP[0.01]).

5.2 Experimental Protocol

Each experiment is performed following the procedure outlined below.

- Extract the initial retrieval results;
- Construct the topical link graph (internal and external links between retrieved XML elements);
- Compute the link score (according to equation 4);
- Normalize the initial and link scores;
- Compute the combined score (DS theory of evidence);

Table 2: iP[0.01] Values & improvement obtained by application of the combined DS theory (Dalian system retrieval results, some topics).

Topic Id	Baseline	Combined mass (DS)	Improvement %	Combined mass (DS) with discounting rate	Improvement %
414	1	0,4204	-57,96	0,4204	-57,96
415	0,5525	0,2333	-57,77	0,2094	-62,10
416	0,0469	0,07258	54,75	0,05871	25,18
417	0,0005	0,0005	0	0,0005	0
419	0,6391	0,7104	11,16	1	56,47
421	0,4175	1	139,52	0,634	51,85
422	0,0386	0,0533	38,08	0,03867	0,18
424	1	1	0	1	0
425	0,8141	1	22,83	1	22,83
426	0,8372	1	19,44	1	19,44
428	1	1	0	1	0
429	0,9479	1	5,49	1	5,49
433	1	0,6188	-38,12	0,7138	-28,62
434	0,9798	0,9812	0,14	0,9812	0,14
436	0,0173	0,0173	0	0,0173	0
473	0,1181	0,1459	23,53	0,1435	21,50
521	0,2107	0,4873	131,27	0,3128	48,45

Table 3: iP[0.01] values obtained by combined DS theory compared to baseline and Topical Pagerank (results over all topics for the three best systems of INEX 2007 Focused task).

	Baseline	Topical Pagerank	Combined mass (DS)	Combined mass (DS) with discounting rate
DALIAN University System	0,5271	0,5470 (+3.78%)	0.5682 (+7.79%)	0.5591 (+6.07%)
Waterloo University System	0.5108	0,5218 (+2.15%)	0.5502 (+7.71%)	0.5484 (+7.36%)
MaxPlanck Institute System	0.5066	0,5072 (+0.11%)	0.5310 (+4.81%)	0.5281 (+4.24%)

- Generate the re-ranked list of XML elements;
- Evaluate the new re-ranked list using INEX evaluation tool.

In our experiment, we have fixed λ parameter of equation 1 to 0.2, which means that a navigational link is 5 times relevant compared to a hierarchical link.

5.3 Experimental Results

From table 2, we note that the proposed approach improves accuracy in most of the topics (i.e. 416, 419, 421, 422, 425, 473, etc.). Thanks to the Demspter–Shafer theory and the link computation approach, the obtained combined results show significant improvement compared to baseline, which conclude to that link evidence can be used as an accurate source of evidence in the XML elements relevance computation process.

We observe that some topics in which improvement is equal to 0 is principally due to the value of the baseline. Our approach gives the same highest value (iP[0.01] = 1), and as a consequence, it confirms the importance of the value of the content evidence. In this case, the link evidence supports the content evidence. However, in the case of topics 417 and 436, the non-improvement is due to the lowest accuracy of the initially retrieved results (topic 417: iP[0.01] = 0.0005). Most of the relevance decreases in table 2 are due to the absence of navigational links between returned XML elements. For instance, topics 414 and 433 which have a baseline iP[0.01] equal to 1, contain only two navigational links. This means that link evidence cannot contribute in the selection of relevant elements, because only few elements will get a high link score.

According to tables 2 and 3, we note that the two variants of combination (with and without discounting rate) improve the retrieval accuracy (for the three systems), and the variant without the discounting rate outperforms the one using the discounting.

Compared to "Topical Pagerank" approach (Mataoui et al., 2010), the two combination DS variants performs better. These results can be interpreted by the use of the "element-element" link type instead of "document-document" link type (used by Topical Pagerank).

Actually, we are experimenting a multitude of discounting rate formulas in order to define an appropriate value allowing best improvements, as well as experimenting our approach using other systems retrieval data.

6 CONCLUSION

We have proposed in this paper an evidential "topic-sensitive" approach based on link path weight for

XML IR. The proposed approach apply a re-ranking process upon initially retrieved XML elements, by evidential combining both XML scores (computed link-based and initial scores). This evidential combination is based on the use of the Demspter–Shafer theory of evidence. Our approach exploits both internal and external links to build specific "element-to-element" links. It introduces a new parameter, called "link weight", in the link score computation. By using the theory of evidence, it combines scores of both bodies of evidence in order to re-rank XML retrieval results. Our proposals are evaluated under the INEX Wikipedia test collection. The results showed improvement compared to baseline and "Topical Pagerank" approach in most of topics. This means that combining link evidence using DS theory with its content evidence outperforms the content-based approach. In future work, we aim to address the behavior of the proposed approach using some Dempster's alternative rules upon multiple systems retrieval results.

ACKNOWLEDGEMENTS

Special thanks to all the people who supported this research, particularly SIG team members of IRIT Institute, France.

REFERENCES

Wikipedia: The free encyclopedia. 2013. http://en.wikipedia.org/.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *The annals of mathematical statistics*, pages 325–339.

Denoyer, L. and Gallinari, P. (2007). The wikipedia xml corpus. In *Comparative Evaluation of XML Information Retrieval Systems*, pages 12–19. Springer.

Dopichaj, P., Skusa, A., and Heß, A. (2009). Stealing anchors to link the wiki. In *Advances in Focused Retrieval*, pages 343–353. Springer.

Fachry, K. N., Kamps, J., Koolen, M., and Zhang, J. (2008). Using and detecting links in wikipedia. In *Focused access to XML documents*, pages 388–403. Springer.

Farahat, A., LoFaro, T., Miller, J. C., Rae, G., and Ward, L. A. (2006). Authority rankings from hits, pagerank, and salsa: Existence, uniqueness, and effect of initialization. *SIAM Journal on Scientific Computing*, 27(4):1181–1201.

Fuhr, N. and Großjohann, K. (2001). Xirql: A query language for information retrieval in xml documents. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 172–180. ACM.

Geva, S., Kamps, J., Lethonen, M., Schenkel, R., Thom, J. A., and Trotman, A. (2010). Overview of the inex 2009 ad hoc track. In *Focused retrieval and evaluation*, pages 4–25. Springer.

Geva, S., Trotman, A., and Tang, L.-X. (2009). Link discovery in the wikipedia. *Pre-Proceedings of INEX 2009*.

Gövert, N. and Kazai, G. (2002). Overview of the initiative for the evaluation of xml retrieval (inex) 2002. In *INEX Workshop*, pages 1–17. Citeseer.

Guo, L., Shao, F., Botev, C., and Shanmugasundaram, J. (2003). Xrank: ranked keyword search over xml documents. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 16–27. ACM.

Itakura, K. Y., Clarke, C. L., Geva, S., Trotman, A., and Huang, W. C. (2011). Topical and structural linkage in wikipedia. In *Advances in Information Retrieval*, pages 460–465. Springer.

Jenkinson, D., Leung, K.-C., and Trotman, A. (2009). Wikisearching and wikilinking. In *Advances in Focused Retrieval*, pages 374–388. Springer.

Kamps, J. and Koolen, M. (2008). The importance of link evidence in wikipedia. In *Advances in Information Retrieval*, pages 270–282. Springer.

Kimelfeld, B., Kovacs, E., Sagiv, Y., and Yahav, D. (2007). Using language models and the hits algorithm for xml retrieval. In *Comparative Evaluation of XML Information Retrieval Systems*, pages 253–260. Springer.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.

Lalmas, M. and Ruthven, I. (1998). Representing and retrieving structured documents using the dempster-shafer theory of evidence: Modelling and evaluation. *Journal of Documentation*, 54(5):529–565.

Lempel, R. and Moran, S. (2001). Salsa: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems (TOIS)*, 19(2):131–160.

Mataoui, M., Mezghiche, M., and Boughanem, M. (2010). Exploiting link evidence to improve xml information retrieval. In *Proceeding de la Confrence Internationale sur l'Extraction et la Gestion des Connaissances Maghreb (EGC-M)*, pages 23–33. ESI.

Pehcevski, J., Vercoustre, A.-M., and Thom, J. A. (2008). Exploiting locality of wikipedia links in entity ranking. In *Advances in Information Retrieval*, pages 258–269. Springer.

Schocken, S. and Hummel, R. A. (1993). On the use of the dempster shafer model in information indexing and retrieval applications. *International Journal of Man-Machine Studies*, 39(5):843–879.

Shafer, G. (1976). *A mathematical theory of evidence*, volume 1. Princeton university press Princeton.

Verbyst, D. and Mulhem, P. (2009). Using collectionlinks and documents as context for inex 2008. In *Advances in focused retrieval*, pages 87–96. Springer.

Zhang, J. and Kamps, J. (2008). Link detection in xml documents: What about repeated links. In *SIGIR 2008 Workshop on Focused Retrieval*, pages 59–66.