A Hybrid Approach for XML Similarity

Joe Tekli[1], Richard Chbeir[1] and Kokou Yetongnon[1]

[1] LE2I Laboratory UMR-CNRS, University of Bourgogne
21078 Dijon Cedex France
joe.tekli@khali.u-bourgogne.fr,
{richard.chbeir, kokou.yetongnon}@u-bourgogne.fr

Abstract. In the past few years, XML has been established as an effective means for information management, and has been widely exploited for complex data representation. Owing to an unparalleled increasing use of the XML standard, developing efficient techniques for comparing XML-based documents becomes essential in information retrieval (IR) research. Various algorithms for comparing hierarchically structured data, e.g. XML documents, have been proposed in the literature. However, to our knowledge, most of them focus exclusively on comparing documents based on structural features, overlooking the semantics involved. In this paper, we integrate IR semantic similarity assessment in an edit distance algorithm, seeking to amend similarity judgments when comparing XML-based documents. Our approach comprises of an original edit distance operation cost model, introducing semantic relatedness of XML element/attribute labels, in traditional edit distance computations. A prototype has been developed to evaluate our model's performance. Experiments yielded notable results.

Keywords: Semi-structured XML-based data, Structural similarity, Information retrieval semantic similarity

1 Introduction

In recent years, W3C's XML (eXtensible Mark-up Language) has been accepted as a major means for efficient data management and exchange. The use of XML ranges over information formatting and storage, database information interchange, data filtering, as well as web services interaction. Due to the ever-increasing web exploitation of XML, an efficient approach to compare XML-based documents becomes crucial in information retrieval (IR).

A range of algorithms for comparing semi-structured data, e.g. XML documents, have been proposed in the literature. All of these approaches focus exclusively on the structure of documents, ignoring the semantics involved. However, in the field of information retrieval (IR), estimating semantic similarity between web pages is of key importance to improving search results [12]. Semantic similarity IR research, as well as the unprecedented abundant use of XML-based documents on the web, incited us to expand existing XML structural similarity so as to take into account semantic relatedness while comparing XML documents.

In order to stress the need for semantic relatedness assessment in XML document comparisons, consider the examples in *Figure 1*.

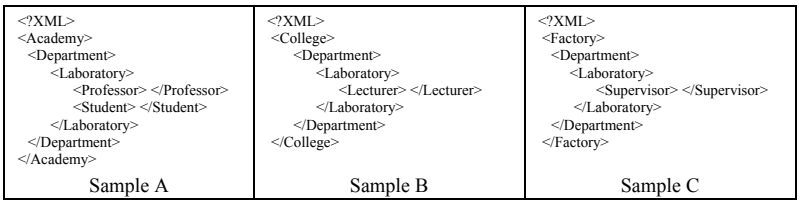

Fig. 1. Examples of XML documents

Using traditional edit distance computations, the same structural similarity value is obtained when document *A* is compared to documents *B* and *C* (Structural similarity computations are detailed in Section 3.2). However, despite having similar structural characteristics, one can obviously recognize that sample document *A* shares more semantic characteristics with document *B* than with *C*. For example, in *Figure 1*, pairs *Academy-College* and *Professor-Lecturer*, from documents *A* and *B*, are semantically similar while *Academy-Factory* and *Professor-Supervisor*, from documents *A* and *C*, are semantically different. It is such *semantic resemblances/differences* that we aim to take into consideration while estimating similarity between XML documents.
In this study, we integrate semantic similarity assessment in a structured XML similarity approach, in order to provide an improved XML similarity measure for comparing heterogeneous XML documents[1]. The remainder of this paper is organized as follows. Section 2 briefly reviews background in both XML structural similarity approaches and IR semantic similarity methods. Section 3 develops our integrated semantic and structure based XML similarity approach. Section 4 presents our prototype and experimental tests, followed by the conclusion in Section 5.

2 Background

2.1 XML Data Model

XML documents represent hierarchically structured information and can be modeled as Ordered Labeled Trees (OLTs) [22]. Nodes in a traditional DOM (Document Object Model) ordered labeled tree represent document elements and are labeled with corresponding element tag names. Element attributes mark the nodes of their containing elements. However, to incorporate attributes in their similarity computations, [14, 24] have considered OLTs with distinct attribute nodes, labeled with corresponding attribute names. Attribute nodes appear as children of their encompassing element nodes, sorted by attribute name, and appearing before all sub-element siblings [14]. In addition, both [14] and [7] agree on disregarding element/attribute values while studying the structural properties of XML documents.

[1] We note by *heterogeneous XML document*, one that doesn't conform to a given grammar (DTD or XML Schema), which is the case of a lot of XML documents found on the web [12].

2.2 XML Structural Similarity

Various methods, for determining structural similarities between hierarchically structured data, particularly XML documents, have been proposed. Most of them derive, in one way or another, the dynamic programming techniques for finding edit distance between strings [10, 20]. In essence, all these approaches aim at finding the cheapest sequence of edit operations that can transform one tree into another. Nevertheless, tree edit distance algorithms can be distinguished by the set of edit operations that are allowed as well as overall complexity and performance levels. Early approaches [23, 19] allow insertion, deletion and relabeling of nodes anywhere in the tree. However, they're relatively greedy in complexity[1]. [4, 6] restrict insertion and deletion operations to leaf nodes and add a move operator that can relocate a sub-tree, as a single edit operation, from one parent to another. However, corresponding algorithms do not guaranty optimal results. Recent work by Chawathe [5] restricts insertion and deletion operations to leaf nodes, and allows the relabeling of nodes anywhere in the tree, while disregarding the move operation. The overall complexity of [5]'s algorithm is of $O(N^2)$. Nierman and Jagadish [14] extend the approach of [5] by adding two new operations: insert tree and delete tree to allow insertion and deletion of whole sub-trees within in an OLT. [14]'s overall complexity simplifies to $O(N^2)$ despite being conceptually more complex than its predecessor. An original structural similarity approach is presented in [7]. It disregards OLTs and utilizes the Fast Fourier Transform to compute similarity between XML documents. Yet, the authors didn't compare their algorithm's optimality to existing edit distance approaches.

2.3 Semantic similarity

Measures of semantic similarity are of key importance in evaluating the effectiveness of web search mechanisms in finding and ranking results [12]. In the fields of Natural Language Processing (NLP) and Information Retrieval (IR), knowledge bases (thesauri, taxonomies and/or ontologies) provide a framework for organizing words (expressions) into a semantic space [8]. Therefore, several methods have proposed to determine semantic similarity between concepts in a knowledge base. They can be categorized as: edge-based approaches and node-based approaches.

The edge-based approach is a natural and straightforward way to evaluate semantic similarity in a knowledge base. [15, 9] estimate the distance between nodes corresponding to the concepts being compared: the shorter the path from one node to another, the more similar they are. Nevertheless, a widely known problem with the edge-based approach is that it often relies on the notion that links in the knowledge base represent uniform distances [16, 8]. In real knowledge bases, the distance covered by a single link can vary with regard to network density, node depth, link type and information content of corresponding nodes [17, 8].

On the other hand, node-based approaches get round the problem of varying link distances. In [16], Resnick puts forward a central node-based method, where the

[1] For instance the approach in [17] has a time complexity $O(|A||B| \text{ depth}(A) \text{ depth}(B))$ when finding the minimum edit distance between two trees A and B ($|A|$ and $|B|$ denote tree cardinalities while depth(A) and depth(B) are the depths of the trees).

semantic similarity between two concepts is approximated by the information content of their most specific common ancestor[1]. Resnick's experiments [16] show that his similarity measure is a better predictor of human word similarity ratings, in comparison with a variant of the edge counting method [15, 9]. Resnick [16] adds that his measure is not sensitive to the problem of varying distances, since it targets the information content of concepts rather than their distances from one another. Improving on Resnick's method [16], Lin [11] presents a formal definition of the *intuitive* notion of similarity, and derives an information content measure from a set of predefined assumptions regarding *commonalities* and *differences*. Lin's experiments [11] show that the latter information content measure yields higher correlation with human judgment in comparison with Resnick's measure [16]. Furthermore, Lin's measure is generalized by Maguitman *et al.* [12] to deal with ontologies of hierarchical (made by IS-A links) and non-hierarchical components (made by cross links of different types), the Lin measure (as most semantic similarity measures) targeting hierarchical structures (taxonomies).
In recent years, there have been a few attempts to integrate semantic and structural similarity in the XML comparison process. The authors in [2, 3, 18] identify the need to support tag similarity (synonyms and stems) instead of tag syntactic equality while comparing XML documents. However, [2, 3, 18] consist of heuristic approaches which disregard the edit distance computations (w.r.t. structure) and only consider the synonymy/stem relations (w.r.t. semantic similarity).

In this study, we aim to combine IR semantic similarity (taking into account the various semantic relations encompassed in the taxonomy/ontology considered in the comparison process) and an edit distance structural similarity algorithm, in order to define a semantic/structural similarity measure for comparing XML documents.

3 Proposal

Our approach consists of an original edit distance operation cost model in which semantic relatedness of XML element/attribute labels is introduced in traditional edit distance computations. In Section 3.1, we present the edit distance process utilized in our study. Then, Section 3.2 develops our integrated semantic/structure based method.

3.1 Structural similarity

Our investigations of the various structural similarity methods proposed in the literature led us to adopt Chawathe's approach [5], his algorithm's performance being recognized and, therefore, further specialized by Nierman and Jagadish [14]. In addition, Chawathe's approach [5] is a direct application of the famous Wagner-Fisher algorithm [20], which optimality was accredited in a broad variety of computational applications [1, 21], updated to take into account tree structures ([20] being originally designed for sequence/string comparisons). Note that integrating semantic similarity assessment in Chawathe's algorithm [5] denotes a straightforward

[1] Note that the *information content* of a concept/class is approximated by estimating the probability of occurrence of the concept/class words in a text corpus

integration of semantic similarity in [14]'s approach, the latter being a *strict generalization* of the former. On the other hand, we adopt Nierman and Jagadish's XML data model [14]. *Figure 2* shows the trees corresponding to the XML document samples presented in *Figure 1*. In fact, we are in agreement with [7, 14]'s decision to disregard element/attribute values while focusing on the structural properties of XML documents adding that, in order to compare element/attribute values, corresponding types should be previously known, which requires prior knowledge of related XML schemas (recall that this study focuses on comparing XML documents lacking DTDs/XML Schemas).

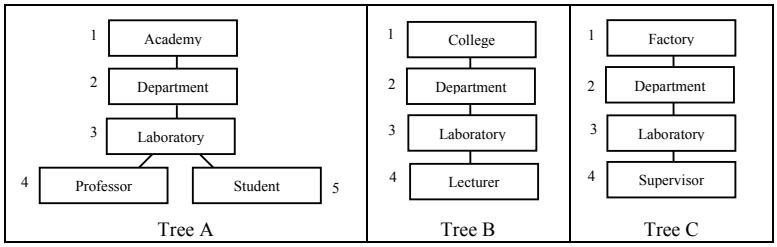

Fig. 2. Motivation example[1]

Chawathe [5] models changes to trees using three basic tree edit operations:

Insertion: *Given a node x of degree 0 (leaf node) and a tree T with root node p having first level sub-trees T_1, …, T_m, Ins(x, i, p, l) is a node insertion applied to T, inserting x as the i^{th} child of p, thus yielding T' with first level sub-trees T_1, … , T_i, x, T_{i+1}, … , T_{m+1}, x bearing l as its label.*

Deletion: *Given a leaf node x and a tree T with root node p, x being the i^{th} child of p, Del(x, p) is a node deletion operation applied to T that yields T' with first level sub-trees T_1, … , T_{i-1}, T_{i+1}, … , T_m.*

Update: *Given a node x in tree T, and given a label l, Upd(x, l) is a node update operation applied to x resulting in T' which is identical to T except that in T', x bears l as its label. The update operation could be also formulated as follows: Upd(x, y) where y.l denotes the new label to be assumed by x.*

By associating costs with each edit operation, Chawathe [5] defines the cost of an edit script (sequence of edit operations) to be the sum of the costs of its component operations. The author in [5] subsequently states the problem of comparing trees: *Given two rooted, labeled, ordered trees A and B, find a minimum cost edit script that transforms A to a tree that is isomorphic[2] to B.* Note that the distance value between two trees *A* and *B* denotes, in a roundabout way, the similarity between them (the

[1] The number next to a node is its preorder rank and serves as node identifier. Please note that there is no correspondence between node identifiers when given two trees to compare. Node correspondence can only be achieved through node labels, taking into consideration their positions in the tree.

[2] Two trees are said to be isomorphic if they are identical except for node identifiers [4].

smaller the distance between *A* and *B*, the more similar they are). Similarity measures based on edit (or metric) distance are generally computed as:

$$\mathrm{Sim}(A, B) = \frac{1}{1 + \mathrm{Dist}(A, B)} \quad \textbf{(1)}$$

On the other hand, a central question in most edit distance approaches is how to choose operation cost values. An intuitive and natural way has been usually used and consists of assigning identical costs to *insertion* and *deletion* operations ($\mathrm{Cost}_{\mathrm{Ins}} = \mathrm{Cost}_{\mathrm{Del}} = 1$), as well as to *update* operations only when the newly assigned label is different from the node's current label ($\mathrm{Cost}_{\mathrm{Upd}}(a, b) = 1$ when $a.l \neq b.l$, otherwise when the labels are the same, $\mathrm{Cost}_{\mathrm{Upd}} = 0$, underlining that no changes are to be made to the label of node *a*). By applying the preceding Intuitive Cost Model (ICM), the edit distance between XML sample trees *A* and *B*, *Dist(A, B)*, following [5] ,would be equal to 3. It is the cost of the following edit script: *Upd(A[1], B[1]), Upd(A[4], B[4]), Del(A[5], A[3]).* The corresponding edit distance computations are shown in *Table 1*. The minimum-cost *ES* contribution to the edit distance computation process is emphasized in *bold* format. Note that an identical edit distance result is attained when comparing sample documents *A* and *C* (*Dist(A, C) = 3*).

Tab. 1. Computing minimum edit distance for XML trees A and B

	0	B[1] (*College*, 0)	B[2] (*Department*, 1)	B[3] (*Laboratory*, 2)	B[4] (*Lecturer*, 3)
0	0	1	2	3	4
A[1] (*Academy*, 0)	1	**1**	2	3	4
A[2] (*Department*, 1)	2	2	**1**	2	3
A[3] (*Laboratory*, 2)	3	3	2	**1**	2
A[4] (*Professor*, 3)	4	4	3	2	**2**
A[5] (*Student*, 3)	5	5	4	3	**3**

As previously mentioned in the introduction, comparing sample documents *A*, *B* and *C*, via strict structural evaluation, yields identical similarity values (the semantics involved being disregarded): *Sim(A,B) = Sim(A, C) = 1/(1+3) = 0.25*
Apparently, intuitive cost models (like the one used previously) do not affect the correctness of Chawathe's structural similarity algorithm [5]. However, they fail to capture the semantics of XML documents. In this study, we propose to complement the structure-based similarity algorithm, developed in [5], with a cost model integrating semantic assessment (IR semantic similarity) in the comparison process.

3.2 Integrated semantic and structure based similarity approach

Apparently, intuitive cost schemes (like the one used previously) do not affect the correctness of the structural similarity algorithm. However, they fail to capture the semantics of XML documents. In this study, we propose to complement Chawathe's edit distance approach [5], with a cost scheme integrating semantic assessment.

3.2.1 Semantic similarity measure

Our investigation of the IR semantic similarity literature led us to consider Lin's similarity measure [11], in our XML comparison process. Lin's measure was proven efficient in evaluating semantic similarity. Its performance and theoretical basis are

recognized and generalized by [12] to deal with hierarchical and non-hierarchical structures. Please bear in mind that our XML similarity process is not sensitive, in its definition, to the semantic similarity measure used. However, choosing a performing measure would yield better similarity judgment. Following Lin [11], the semantic similarity between two words (expressions) can be computed as follows:

$$\mathrm{Sim}_{\mathrm{Sem}}(w_1, w_2) = \mathrm{Sim}_{\mathrm{Sem}}(c_1, c_2) = \frac{2 \log p(c_0)}{\log p(c_1) + \log p(c_2)} \quad (2)$$

- c_1 and c_2 are concepts, in a knowledge base of hierarchical structure (taxonomy), subsuming words w_1 and w_2 respectively
- c_0 is the most specific common ancestor of concepts c_1 and c_2
- *p(c)* denotes the occurrence probability of words corresponding to concept *c*. It can be computed as the relative frequency: p(c) = freq(c) / N
  - $\mathrm{freq}(c) = \sum_{w \in \mathrm{words}(c)} \mathrm{count}(w)$: sum of the number of occurrences, of words subsumed by *c*, in a given corpus
  - N: total number of words encountered in the corpus

In information theory, the *information content* of a class or concept c is measured by the negative log likelihood *-log p(c)* [16, 12]. While comparing two concepts c_1 and c_2, Lin's measure takes into account each concept's *information content* ($-\log p(c_1) + -\log p(c_2)$), as well as the *information content* of their most specific common ancestor ($-\log p(c_0)$), in a way to increase with commonality (*information content* of c_0) and decrease with difference (*information content* of c_1 and c_2) [11].

3.2.2 Label semantic similarity cost

To take into account semantic similarity in XML comparisons, while utilizing the edit distance algorithm, we propose to vary operation costs according to the semantics of concerned nodes. While comparing XML sample documents *A-B* and *A-C* for example, the similarity evaluation process should realize that elements *Academy-College* have higher semantic similarity than *Academy-Factory*. Likewise, *Professor-Lecturer* have higher semantic similarity than *Professor-Supervisor*. Therefore, overall similarity *Sim(A, B)* should be of greater value vis-à-vis *Sim(A, C)*. Such semantic relatedness would be taken into consideration by varying operation costs as follows:

$$\mathbf{Cost_{Sem\_Upd}(x, y) = 1 - Sim_{Sem}(x.l, y.l)} \quad (3)$$

The more the *initial* and the *replacing* node labels (*x.l* and *y.l* respectively) are semantically similar, the lesser the update operation cost, which transitively yields a lesser minimum cost *ES* (higher similarity value). When labels are identical, semantic similarity is of maximum value, $Sim_{Sem}(x.l, y.l) = 1$, yielding $Cost_{Upd}(x, y) = 0$ (no changes to be made). When labels are completely different, semantic similarity is of minimum value, $Sim_{Sem}(x.l, y.l) = 0$, which brings us to $Cost_{Upd}(x, y) = 1$.
Following the same logic, we consider varying insertion and deletion costs.

$$\mathbf{Cost_{Sem\_Ins}(x, i, p, l) = 1 - Sim_{Sem}(l, p.l)} \quad (4)$$

$$\mathbf{Cost_{Sem\_Del}(x, p) = 1 - Sim_{Sem}(x.l, p.l)} \quad (5)$$

While inserting or deleting a node from an XML document, we evaluate semantic relatedness between the inserted/deleted node's label and the label of its ancestor in the document tree. The more an inserted/deleted node label is semantically similar to its ancestor node label, the lesser the insertion/deletion operation cost, which transitively yields a lesser cost *ES* (higher similarity value). When labels are identical or completely different, insertion/deletion costs would be equal to 0 or 1, respectively[1] (as with the update operation). Such semantic assessments would reflect semantic relatedness between inserted/deleted nodes and their context, in the XML document, affecting overall similarity accordingly.
Furthermore, our investigation of semantic similarity, in XML documents, led us to consider varying operation costs with respect to node depth.

3.2.3 Node depth cost

Node depth consideration in XML document comparison is not original in the literature. Zhang *et al.* [24] have already addressed the issue. Following [24], editing the root node of an XML tree would yield significantly greater change than editing a leaf node. Notionally, as one descends in the XML tree hierarchy, information becomes increasingly specific, consisting of finer and finer details, its affect on the whole document tree decreasing accordingly. For example, consider the XML sample tree *A* in *Figure 2*. Editing node *A[1]* (*A[1].l = Academy*) by changing its label to *Hospital*, would semantically affect tree *A* a lot more than deleting node *A[4]* (*A[4].l = Professor*), changing *A*'s whole semantic context. Therefore, it would be relevant to vary operation costs following node depths, assuming that operations near the root node have higher impact than operations further down the hierarchy. The following formula, adapted from [24], could be used for that matter:

$$\mathbf{Cost_{Depth\_Op}(x) = \frac{1}{(1 + x.d)}} \quad \mathbf{(6)}$$

- *Op* is an insert, delete or update operation
- *x.d* is the depth of the node considered for insertion, deletion or updating

The preceding formula assigns unit cost (=1, maximum cost) when the root node is considered and yields decreasing costs when moving downward in the hierarchy.

3.2.4 Semantic cost model (SCM)

In order to take into account semantic meaning while comparing XML documents, we propose to complement Chawathe's edit distance algorithm [5], with the following cost model:

---

[1] In this study, we assume that an XML node and its ancestor cannot have identical labels. However, such cases this will be addressed in future work.

$$\mathbf{Cost_{Op}(x, y) = Cost_{Sem\_Op}(x, y) \times Cost_{Depth\_Op}(x)} \quad (7)$$

– *Op* designates an insertion, deletion or update operation

The results attained by applying the semantic cost model to compare sample XML documents *A*, *B* and *C* are shown in tables *2* and *3*. Note that semantic similarity values between node labels were estimated using Lin's measure [11] (applied on an independently constructed corpus and taxonomy), and are reported in *Table 4.*

Tab. 2. Computing edit distance, via our SCM, for XML sample trees A and B

	0	B[1] (*College*, 0)	B[2] (*Department*, 1)	B[3] (*Laboratory*, 2)	B[4] (Lecturer, 3)
0	0	1	1.5	1.8333	2.0833
A[1] (*Academy*, 0)	1	**0.1148**	0.5365	0.8205	0.9824
A[2] (*Department*, 1)	1.4217	0.5365	**0.1148**	0.1425	0.3413
A[3] (*Laboratory*, 2)	1.4494	0.5642	0.1425	**0.1148**	0.3172
A[4] (*Professor*, 3)	1.651	1.7658	0.3441	0.3164	**0.163**
A[5] (*Student*, 3)	1.8466	1.9614	0.5397	0.512	**0.3586**

Tab. 3. Computing edit distance, via our SCM, for XML trees A and C

	0	B[1] (*Factory*, 0)	B[2] (*Department*, 1)	B[3] (*Laboratory*, 2)	B[4] (*Supervisor*, 3)
0	0	1	1.5	1.8333	2.0833
A[1] (*Academy*, 0)	1	**0.8581**	1.2798	1.5638	1.7813
A[2] (*Department*, 1)	1.4217	1.2798	**0.8581**	0.8858	1.0894
A[3] (*Laboratory*, 2)	1.4494	1.3075	0.8858	**0.8581**	1.0647
A[4] (*Professor*, 3)	1.651	1.5091	1.0874	1.0597	**1.0673**
A[5] (*Student*, 3)	1.8466	1.7047	1.283	1.2553	**1.2628**

Tab. 4. Word semantic similarities, computed following Lin's measure [11]

Word pairs		Sim_{Lin}	Word pairs		Sim_{Lin}
Academy	College	0.8851	Department	Professor	0.2083
Academy	Department	0.1566	Department	Student	0.2367
Academy	Factory	0.1419	Department	Supervisor	0.1857
Academy	Laboratory	0.1481	Factory	Laboratory	0.1963
Academy	Lecturer	0.3521	Factory	Lecturer	0.1803
Academy	Professor	0.3563	Factory	Professor	0.1831
Academy	Student	0.3876	Factory	Student	0.2047
Academy	Supervisor	0.1297	Factory	Supervisor	0.4672
College	Department	0.1566	Laboratory	Lecturer	0.1903
College	Factory	0.1419	Laboratory	Professor	0.1935
College	Laboratory	0.1481	Laboratory	Student	0.2177
College	Lecturer	0.3521	Laboratory	Supervisor	0.1738
College	Professor	0.3563	Lecturer	Professor	0.807
College	Student	0.3876	Lecturer	Student	0.5028
College	Supervisor	0.1297	Lecturer	Supervisor	0.1611
Department	Factory	0.2117	Professor	Student	0.5114
Department	Laboratory	0.9169	Professor	Supervisor	0.1633
Department	Lecturer	0.2047	Student	Supervisor	0.1803

By applying our SCM, the edit distances computed between pairs *A*-*B* and *A*-*C* are no longer identical (in comparison with the intuitive cost scheme):

– $Sim_{SCM}(A, B) = 1 / 1 + Dist_{SCM}(A, B) = 0.7361$; having $Dist_{SCM}(A, B) = 0.3586$
– $Sim_{SCM}(A, C) = 1 / 1 + Dist_{SCM}(A, C) = 0.4418$; having $Dist_{SCM}(A, C) = 1.2628$

Considering semantic relatedness, in the comparison process, reflects the fact that sample documents *A* and *B* are more similar than *A* and *C* ($Sim_{SCM}(A, B) > Sim_{SCM}(A,C)$), in spite of sharing identical structural similarities.
Our SCM, used with a structure-based (edit distance) similarity algorithm, seems to capture semantic meaning effectively, while comparing XML documents.

4 Experimental Evaluation

4.1 Prototype

To validate our approach, we have implemented (using C#) a prototype, entitled "XML SS Similarity", encompassing a v*alidation component*, verifying the integrity of XML documents, and an *edit distance component* undertaking XML similarity computations following the algorithm adopted in our study. In addition, a *synthetic XML data generator* was also implemented in order to produce sets of XML documents based on given DTDs. The synthetic XML generator accepts as input: a DTD document and a MaxRepeats[1] value designating the maximum number of times a node will appear as child of its parent (when * or + options are encountered in the DTD). Furthermore, a *taxonomic analyzer* was also introduced so as to compute semantic similarity values between words (expressions) in a given taxonomy. Our *taxonomic analyzer* accepts as input a hierarchical taxonomy *HT* and corresponding corpus-based word occurrences. Consequently, concept frequencies are computed and, thereafter, used to compute semantic similarity between pairs of nodes in *HT*.

4.2 Experimental results

Various experiments were conducted in order to test the performance of our integrated similarity model. Real and generated (synthetic) XML documents as well as a number of hierarchical taxonomies where considered. In the following, we present the results attained using synthetic XML documents (cf. *Figure 3*) and a WordNet[2]-based hierarchical taxonomy comprising of 677 nodes.

```
<!DOCTYPE DTD1 [
<!ELEMENT Academy (Administrative unit+)>
    <!ELEMENT Administrative unit (Branch?)>
        <!ELEMENT Branch (Educator?, Student+)>
            <!ELEMENT Educator (#PCDATA)>
            <!ELEMENT Student (#PCDATA)>  ]>
```

```
<!DOCTYPE DTD2 [
<!ELEMENT School (Administrative unit+)>
    <!ELEMENT Administrative unit (Section?)>
        <!ELEMENT Section (Educator?, Scholar*)>
            <!ELEMENT Educator (#PCDATA)>
            <!ELEMENT Scholar (#PCDATA)>  ]>
```

```
<!DOCTYPE DTD3 [
<!ELEMENT Government (Administrative unit+)>
    <!ELEMENT Administrative unit (Section?)>
        <!ELEMENT Section (Professional?, Worker+)>
            <!ELEMENT Professional (#PCDATA)>
            <!ELEMENT Worker (#PCDATA)>  ]>
```

Fig. 3. DTDs inducing sets of synthesized XML documents

In this experiment, we evaluate our model's efficiency by assessing similarity results to the *a priori* know DTDs (inducing document sets). Therefore, average inter-set and intra-set[3] similarities are depicted in a matrix where element (i, j) underscores the

[1] A greater MaxRepeats value underlines a greater variability when + and * are encountered.

[2] WordNet is an online lexical reference system (taxonomy), developed by a group of researchers at Princeton University NJ USA, where nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing a lexical concept [13].

[3] Intra-set average similarities are computed between documents of the same set S_i, reported as (i, i) values in the similarity matrix. Remaining (i, j) values correspond to intra-set average similarities, computed between documents belonging to sets S_i and S_j

average similarity value corresponding to every pair of distinct documents belonging to sets S_i (DTD_i) and S_j (DTD_j). Results[1] are reported in two matrixes, corresponding to the intuitive cost model and to our SCM:

Tab. 5. Intuitive cost model (ICM)

	S1	S2	S3
S1	0.5885	0.0951	0.0982
S2	0.0951	0.1515	0.0945
S3	0.0982	0.0945	0.4110

Tab. 6. Semantic Cost Model (SCM)

	S1	S2	S3
S1	0.8877	0.3403	0.3325
S2	0.3403	0.4392	0.3363
S3	0.3325	0.3363	0.6400

First of all, results show that our SCM produces relatively higher similarity values, underlining similarities (of semantic nature) that were undetected using the ICM. On the other hand, a straight distinction between documents belonging to a set and others outside that set is attained with our SCM, as with the ICM (comparing highlighted values, in tables *5* and *6*, to remaining values). Furthermore, our SCM captures semantic affinities between documents belonging to different sets. For instance, a relatively higher average similarity degree is attained between sets S_1 and S_2 (*0.3403* - DTDs *1* and *2* revealing similar semantic content), in comparison with S_1 and S_3 (*0.3325*), the average similarity value between S_1/S_2 (ICM: *0.0951*) being lesser than that of S_1/S_3 (ICM: *0.0982*) using the ICM.
Nonetheless, the improved structural/semantic similarity results aren't attained without affecting overall time complexity. In brief, our complexity tests show that the time to compute similarity grows in an almost perfect linear fashion, when using the classic ICM (complexity of Chawathe's classic edit distance approach [5]). However, when introducing our SCM, it incrementally shifts towards a polynomial (quadratic) function, following the growing number of taxonomic nodes involved (detailed complexity results are omitted due to spatial constraints).

5 Conclusion and future work

In this paper, we proposed an integrated semantic and structure based XML similarity approach, taking into account the semantic meaning of XML element/attribute labels in XML document comparison. To our knowledge, this is the first attempt to combine edit distance structural similarity computations with IR semantic similarity assessment, in an XML (structured data) context. Experimental results confirmed the positive impact of semantic meaning on XML similarity values, and reflected its heavy impact regarding complexity.

Future directions include exploiting semantic similarity to compare, not only the structure of XML documents (element/attribute labels), but also their information content (element/attribute values). In such a framework, XML Schemas seem unsurpassable, underlining element/attribute data types, required to compare corresponding element/attribute values. The semantic complexity problem will also be tackled in upcoming studies.

[1] In this experiment, we generated synthetic XML documents with MaxRepeats = 10

References

1. Aho A., Hirschberg D., and Ullman J., Bounds on the Complexity of the Longest Common Subsequence Problem. *Journal of the Association for Computing Machinery*, 23(1):1-12, January 1976.
2. Bertino E., Guerrini G., Mesiti M., Rivara I. and Tavella C., Measuring the Structural Similarity among XML Documents and DTDs, Technical Report, University of Genova, 2002, http://www.disi.unige.it/person/MesitiM.
3. Bertino E., Guerrini G., Mesiti M., A Matching Algorithm for Measuring the Structural Similarity between an XML Documents and a DTD and its Applications, *Elsevier Computer Science*, 29 (23-46), 2004.
4. Chawathe S., Rajaraman A., Garcia-Molina H., and Widom J., Change Detection in Hierarchically Structured Information. In *Proc. of the ACM Int. Conf. on Management of Data (SIGMOD)*, Montreal, Quebec, Canada, 1996.
5. Chawathe S., Comparing Hierarchical Data in External Memory. In *Proceedings of the Twenty-fth Int. Conf. on Very Large Data Bases,* p. 90-101, 1999.
6. Cobéna G., Abiteboul S. and Marian A., Detecting Changes in XML Documents. In *Proc. of the IEEE Int. Conf. on Data Engineering*, p. 41-52, 2002.
7. Flesca S., Manco G., Masciari E., Pontieri L., and Pugliese A., Detecting Structural Similarities Between XML Documents. In *Proc. of WebDB 2002*, 2002.
8. Jiang J. and Conrath D., Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. *In Proc. of the Int. Conf. on Research in Computational Linguistics*, 1997.
9. Lee J.H., Kim M.H. and Lee Y.J., Information Retrieval Based on Conceptual Distance in IS-A Hierarchies. *Journal of Documentation*, 49(2):188-207, 1993.
10. Levenshtein V., Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Sov. Phys. Dokl*., 6:707-710, 1966.
11. Lin D., An Information-Theoretic Definition of Similarity. *In Proceedings of the 15th Int. Conf. on Machine Learning*, 296-304, Morgan Kaufmann Pub. Inc., 1998.
12. Maguitman A. G., Menczer F., Roinestad H. and Vespignani A., Algorithmic Detection of Semantic Similarity. In *Proc. of the 14th Int. WWW Conference*, 107-116, Japan, 2005.
13. Miller G., WordNet: An On-Line Lexical Database. *Int. Journal of Lexicography*, 1990.
14. Nierman A. and Jagadish H. V., Evaluating structural similarity in XML documents. In *Proc. of the 5th Int. Workshop on the Web and Databases,* 2002.
15. Rada R., Mili H., Bicknell E. and Blettner M., Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19:17-30, 1989.
16. Resnik P., Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proc. of the 14th IJCA-95*, Vol. 1, 448-453, Montreal, Canada, 1995.
17. Richardson R. and Smeaton A.F., Using WordNet in a Knowledge-based approach to information retrieval. In *Proc. of the 17th Colloquium on Information Retrieval*, 1995.
18. Sanz I., Mesiti M., Guerrini G. and Berlanga Lavori R., Approximate Subtree Identification in Heterogeneous XML Documents Collections. *XSym*, 192-206, 2005.
19. Shasha D. and Zhang K., Approximate Tree Pattern Matching. In *Pattern Matching in Strings, Trees and Arrays*, chapter 14, Oxford University Press, 1995.
20. Wagner J. and Fisher M., The String-to-String correction problem. *Journal of the Association of Computing Machinery*, 21(1):168-173, 1974.
21. Wong C. and Chandra A., Bounds for the String Editing Problem. Journal of the Association for Computing Machinery, 23(1):13-16, January 1976.
22. WWW Consortium, The Document Object Model, http://www.w3.org/DOM.
23. Zhang K. and Shasha D., Simple Fast Algorithms for the Editing Distance Between Trees and Related Problems. *SIAM Journal of Computing*, 18(6):1245-1262, December 1989.
24. Zhang Z., Li R., Cao S. and Zhu Y., Similarity Metric in XML documents. *Knowledge Management and Experience Management Workshop*, 2003.