Semantic Structure for XML Documents: Structuring and Pruning

Salma Ben Mefteh[1,2], Kaïs Khrouf [1], Jamel Feki[1], Chantal Soulé-Dupuy[2]
[1]Laboratory Mir@cl
University of Sfax
Sfax, Tunisia
[2]Laboratory IRIT
University of Toulouse I Capitole
Toulouse, France
{Salma.BenMefteh, Jamel.Feki}@fsegs.Rnu.tn, Khrouf.Kais@isecs.Rnu.tn, Chantal.Soule-Dupuy@ut-capitole.fr

ABSTRACT: *The nature of information sources together with the multiple document description alternatives offered by these sources are permanently changing. This evolvement was accelerated by the expansion of the Web and, therefore has increased the volume of information available as XML documents. These latter are described through semantic-less tags (e.g., Content, Section, Paragraph). Despite the real and urgent need for a semantically exploitation of these XML documents, there are too few research issues tackling the semantic aspects of XML documents. Therefore, the development of automated tools able to determine semantic structures for XML documents is becoming a necessity and a challenge. In this paper, we propose a novel automated approach to create semantic structures for XML documents in order to perform semantic OLAP analyses. The software prototype developed to support this approach is also described.*

Keywords: Component, XML Document, Semantic Structure, Ontology

Received: 28 November 2012, Revised 30 December 2012, Accepted 3 January 2013

1. Introduction

The development of the Internet has widely increased the number of documents and data volumes available and exchanged through the Web. Face to this mass of documents, XML has become the standard format for documents. Thus, a huge number of increasingly pertinent documents become more and more available in this format. XML documents can be classified into two types: *data-centric* XML documents and *document-centric* XML documents.

Data-centric XML documents contain short and precise data; in fact, they are very similar to identifiers in relational databases. This type of document is mostly used by applications exchanging information (i.e., transactional data). In such documents, tags precisely describe the content, and then provide the necessary semantic for the comprehension of information contained within the document (e.g., Product, Customer, Quantity, Price are meaningful tags).

Whereas, document-centric XML documents are text-rich documents; they constitute the electronic version of traditional paper documents (e.g.,scientific articles, internal reports, e-books). Tags used for such documents (e.g., Content, Section, Paragraph) typically describe their logical structure but not their semantics. This is a major drawback in decision support system relying on OLAP (On-Line Analytical Processing) analyses where semantic information are necessary.

Consequently, the proposal of an approach dealing with semantics of documents and the development of an automated software tool supporting it become a necessity and a true challenge. To this end, we propose a novel approach for automatic extraction of semantic structures of document-centric XML documents. It is intended for OLAP analyses on documents. In [3], the authors define the semantic structure as "*a set of semantic tags representing concepts associated through relationships*". In our context, we consider that the semantic structure of an XML document is a hierarchical structure of concepts that enhances the logical structure; mainly, it synthetically describes the semantics of the document's textual components.

The remainder of this paper is organized as follows. Section 2 describes related works dealing with the semantic structure of documents. Section 3 presents our approach for determining the semantic structure. Section 4 overviews some results issued from the software prototype we have implemented. Finally, Section 5 summarizes the paper and enumerates interesting perspectives.

2. Related Work

The majority of works that addressed the semantic of documents (texts and XML documents) was interested in the semantic representation of the content of these documents.

In information retrieval domain, the semantic aspect was discussed to improve the number of restituted documents compared to a given user query. The authors of [11] proposed a method for evaluating similarity between the terms of an XML document. Their method was based on the vector model [9]; it is implemented within a system of semantic indexing of XML documents. The idea consists to replace the terms issued from a user query by their associated concepts in order to deliver the most appropriate responses. The major drawback of this work is the use of a specific ontology of concepts related to a corpus. In the same spirit, and in addition to indexing the textual content of documents, the authors of [6] index also the structures of documents by using the concepts of the WordNet lexical database. However, in that case only the most common terms of documents or queries are referenced by concepts.

In [2], it has been proposed a model for semantic representation of documents and queries based on a semantic network (i.e., a set of nodes connected by arcs, where nodes represent concepts and arcs represent *is-a* relationships between these concepts). However, this work claims that the semantic indexation (i.e., assigning a set of concepts to a document) does not improve the query results except when it is combined with a classic indexation method based on keywords.

On the other hand, in [5] and [11]the authors were interested with the classification of documents.

The authors of [5] propose a classification approach of documents by semantic enrichment. Their approach consists in enriching short texts (generally articles of press) by using domain ontologies. They propose three types of enrichment: (1) enrichment by generalization (e.g., the current events speaking about "*Michaël Jackson*" will be enriched by the "*Rock*" concept), (2) enrichment by specialization ("*Sport*" can be enriched by the media sports), and (3) enrichment by categorization (Addition of meta-data as Author, Year…). Nevertheless, the classification is supervised: the classes of documents have to be known a priori.

In [11], the authors presented a classification method for texts; it is based on statistical and semantic techniques. This classification is realized in three steps: (1) Construction, using Wordnet, of two vectors: one for the document terms and another for each activity domain (e.g., Medicine, Cryptography), (2) Calculation of score between the vector of documents and all domain vectors, and (3) Assignment of the domain vector having the highest score to the appropriate document. However, this work uses WordNet as ontology which is too general so that the accuracy of the result is weak.

In the literature of document content structuring, few works were interested in the semantic structures of documents [3] [10]. Thus, [3] proposes an approach for accessing documents (Ph.D thesis) by semantic contents. They proposed a model of documents which is based on the use of new metadata (called "*semantic tags*"), in order to refine the search and better satisfy the user requests. However, the proposed approach is intended for one single type of documents: PhD thesis. At his side, [10] proposes to semantically enrich the tags of XML documents. To do so, the authors consider that each path in the XML document represents a network and each tag of this path constitutes a layer (the set of meanings of the tag, taken from WordNet). A next step calculates a similarity measure between layers; it is to determine the best path through the network. Nevertheless, the use of WordNet may cause difficulties in selecting the most appropriate meaning for a given tag, especially for poly semic words.

As a complement to the work dealing with the semantic content, we address in this paper the problem of *how to semantically structure XML documents*. More precisely, we propose a novel automatic approach for the extraction of semantic structures from XML documents. This approach is based on logical structure and content of documents, unlike the proposal in [10] which focused only on tag names. Our work enables visualizing and querying XML documents according to logical and/or semantic viewpoints.

3. ProposedApproach

In [7], we have proposed an approach for the classification and the multidimensional analysis of documents. This approach gathers identical or similar logical structures of XML documents into generic structures (the proposed method for comparing logical structures is described in [4]).

The multidimensional analysis described in [7] was validated for data-centric XML documents. As our objective is to extend this work to document-centric XML documents (reports, scientific papers, news...) then we propose to derive for an XML document an additional structure that reflects its semantics. This derivation relies simultaneously on the logical structure of the document and on its contents. This represents the core topic of this paper. Figure 1 depicts an introductive example of a logical structure and a semantic structure for an XML document.

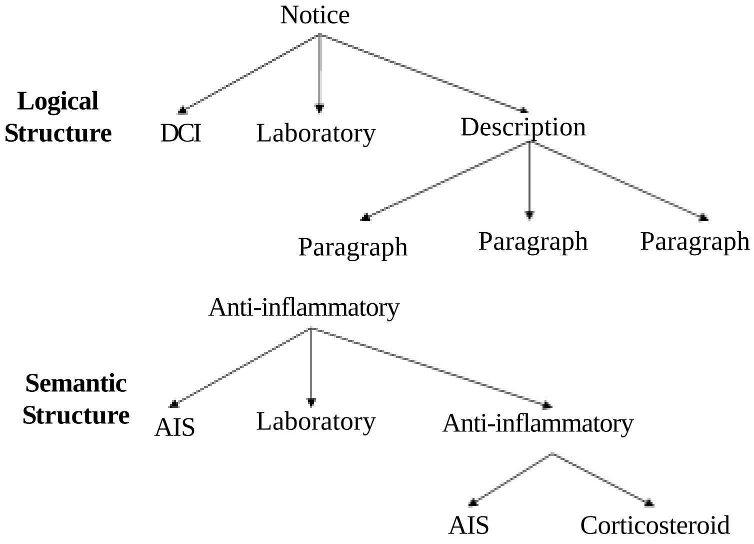


Figure 1. Example of a logical structure and its associated semantic structure

Figure 2 illustrates our proposed approach for the automatic construction of semantic structures.

This approach includes 5 phases:

- **Terms extraction:** Extracts significant keywords for leaf elements of the logical structure of documents (textual fragments).
- **Ontology selection:** Identifies, among a set of domain ontologies, one ontology that better matches the domain of the document.
- **Assigning concepts to leaf elements:** assigns to every leaf element of a document the most significant concept from the ontology selected in the previous phase.
- **Propagation of concepts:** attributes a concept to each non-leaf element of the logical structure relying on the concepts of its child elements.
- **Pruning the semantic structure:** keeps the elements representing metadata, removes elements without corresponding concept

and finally replaces some elements assigned to the same concept by a single element.

Note that for the first phase we have used information retrieval techniques [1]. Phases 2 to 5 are detailed hereafter.

3.1 Ontology selection

The objective of this phase is to create a first outline of the semantic structure for every XML document, essentially by exploring multiple domain ontologies. Specifically, we should determine and associate a single ontology to each document. To do so, we assume that the ontology concepts are weighted by an expert. These weights reflect the importance of concepts in the ontology for the end-user.

So then, for every domain ontology, we firstly calculate the weight of every concept C_i relative to each leaf element E_j of document *d* (cf. Formula 1).

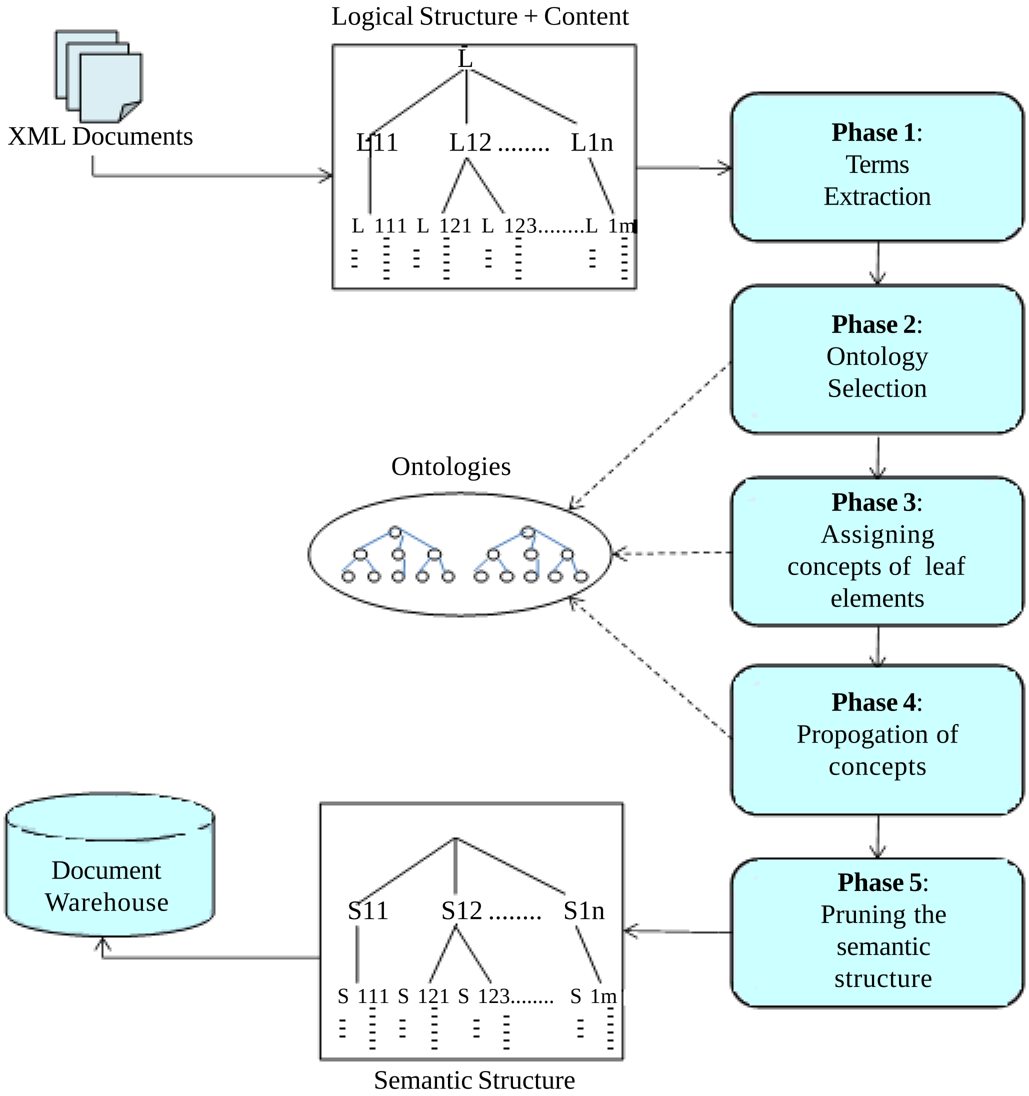

Logical Structure + Content
XML Documents
L
L11
L12 ....
L1n
L 111 L 121 L 123........L 1m
Phase 1: Terms Extraction
Phase 2: Ontology Selection
Ontologies
Phase 3: Assigning concepts of leaf elements
Phase 4: Propogation of concepts
Document Warehouse
S11
S12 ....
S1n
S 111 S 121 S 123........ S 1m
Phase 5: Pruning the semantic structure
Semantic Structure

Figure 2. An approach to determine a semantic structure for a document-centric XML document

$$Weight\,(C_i^{E_j}) = \frac{\left|C_i^{E_j}\right|}{\left|C_i^{d}\right|} * Weight\,(C_i^{O_k})\ \forall j,\ E_j \in d \qquad (1)$$

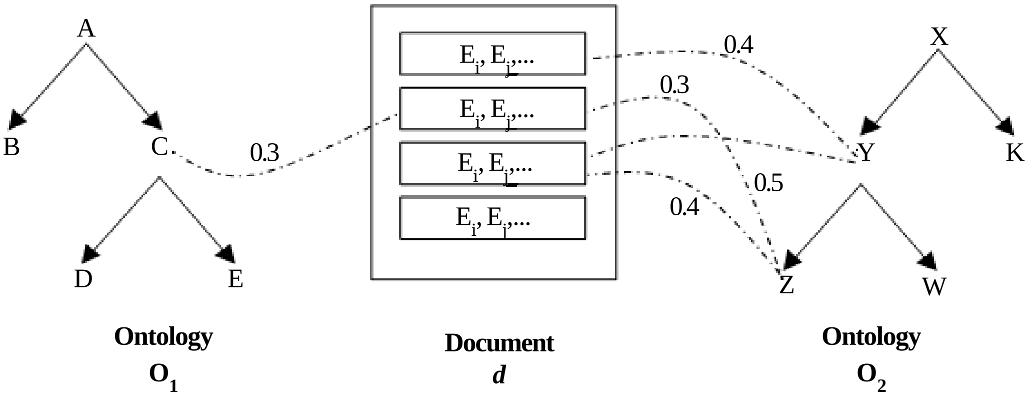


Figure 3.Weights of concepts calculated for each document element

where:

- $\left|C_i^{E_j}\right|$ is the number of occurrences of concept C_i in element E_j,

- $\left|C_i^{d}\right|$ is the number of occurrences of C_i in document d, and

- $Weight\,(C_i^{O_k})$ is the weight of the concept C_i in ontology O_K.

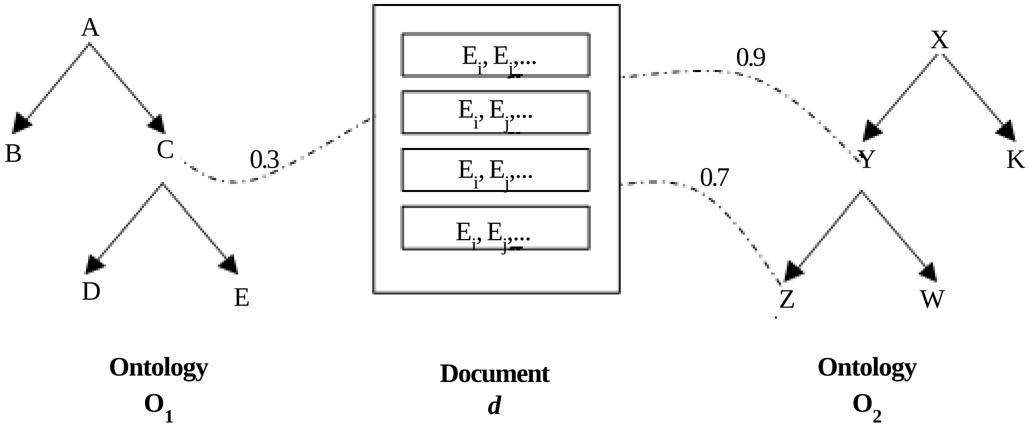


Figure 4.Weights of concepts calculated for the whole document

After that, we calculate the weight of each concept C_i in document d according to Formula 1. This weight is equal to the sum of weights of C_i in all elements of d.

$$Weight\,(C_i^{\,d}) = \sum_{j=1}^{N} Weight\,(C_i^{E_j})\ \forall j,\ E_j \in d \qquad (2)$$

where: N is the number of elements in document d.

To choose the most appropriate ontology for a document, we calculate the sum of weights of the concepts belonging to the same ontology (cf. Formula 3).

$$Weight\,(O_k^{\,d}) = \sum_{i=1}^{|O_k|} Weight\,(C_i^{\,d}) \qquad (3)$$

where: $|O_k|$ is the number of concepts in ontologyO_k

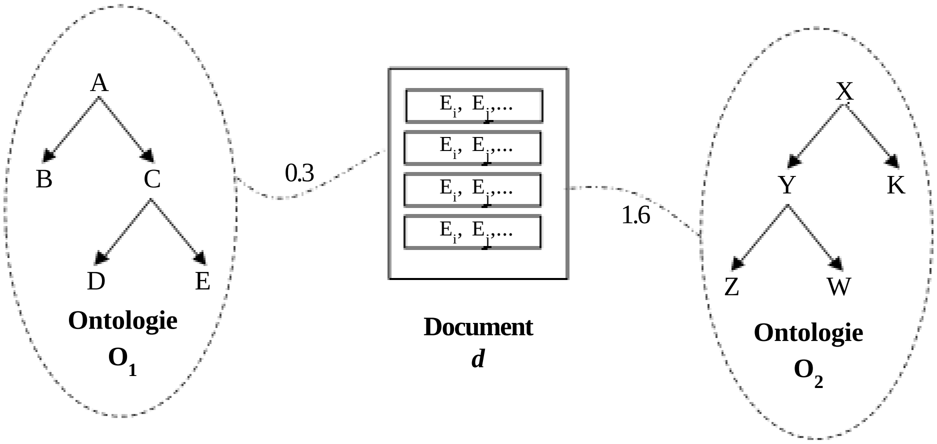


Figure 5. Weights of each ontologyfor the whole document

Finally, for document *d*, we select the ontology having the highest weight computed according to Formula 3. In our example, the ontology O_2 has the highest weight 1.6; therefore it will be elected for the document *d*.

In the remainder of this paper, we assume that the document having its logical structure shown in Figure 6 is assigned to the Data Warehouse ontology of Figure 7.

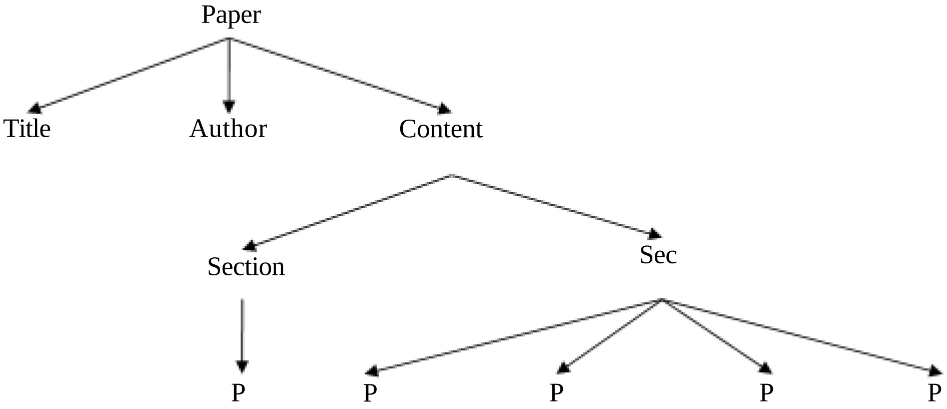


Figure 6.The logical structure of document *d*

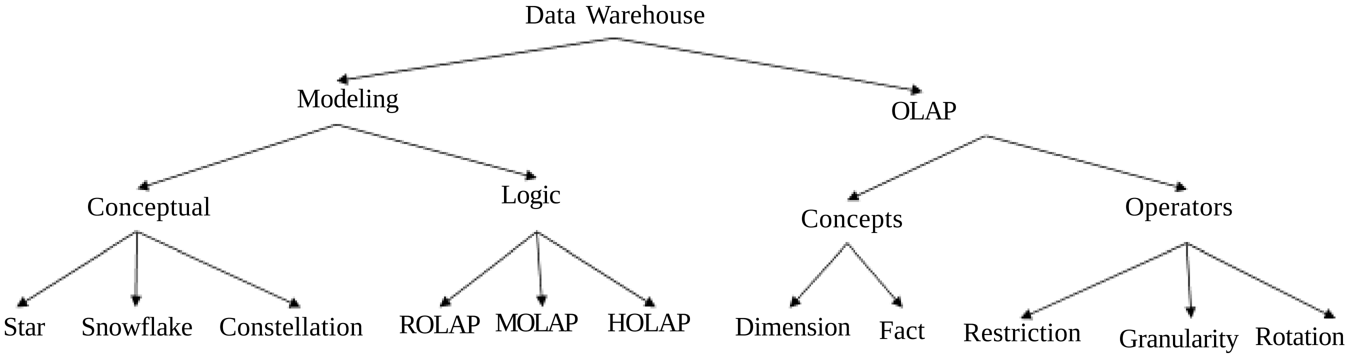


Figure 7. The Data Warehouse ontology

3.2 Assigning concepts to leaf elements

As the objective of this phase is to assign one representative concept to every leaf element of the document structure, we use weights of concepts calculated by Formula 1.

For a leaf element E_k, four different cases may occur:

- **Case 1:** No associated concept for E_k as for the meta-data elements *Author*, *Publisher* and *Year*. Thus, the *Null* concept will be assigned to E_k.

- **Case 2:** A single concept is determined for E_k; it will be assumed as the unique representative concept.

- **Case 3:** Several concepts are identified for E_k, all of them belonging to the same hierarchy within the selected ontology. In this case, we envisage two situations:

✓ If the weights calculated for these concepts are almost similar (i.e., the difference between each pair of weights is less than a given threshold 0.1) then we assign to E_k the most specific one among these concepts in the hierarchy; i.e., concept which has the lowest hierarchical level.

✓ Otherwise, we assign to E_k the concept having the highest weight, independently of its position in the hierarchy.

- **Case 4:** Several concepts are identified for E_k and they belong to distinct hierarchies in the ontology. In this situation, we assign the concept having the highest weightto E_k.

At the end of this phase, every leaf element is associated with a single concept (even the *Null*) of the selected ontology.

For our running example, the result of this step on document *d* is illustrated in Figure 8.

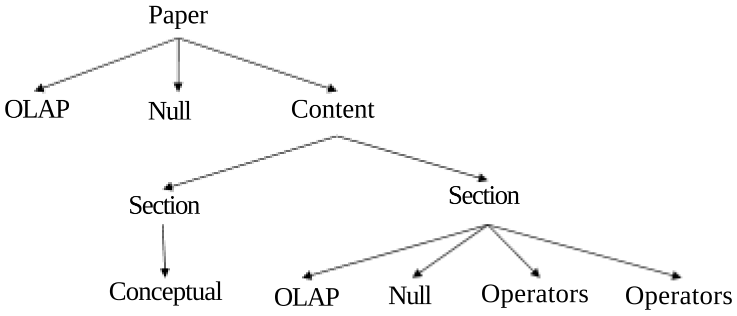


Figure 8. Concepts assigned to leaf elements (in bold)

3.3 Propagation of concepts

The propagation of concepts is a derivation process of concepts; it aims to derive a concept for node *n* at level *i* of the semantic structure, under construction, by rolling up the immediate sub-concepts of *n* in the ontology. The objective of the propagation is to assign derived concepts to non-leaf elements of the specific structure from leaf elements until we reach the root element. We accomplish this propagation through a set of three rules defined hereafter:

- **Rule P1:** In the semantic structure, a father node having a single valued child node (i.e., one concept) will be assigned the same concept as its child.

- **Rule P2:** In the semantic structure, if a father node has several sub-elements whose concepts belong to the same hierarchy of the ontology, then this father will have the most generic concept of the concepts associated with his children.

- **Rule P3:** If a father node *n* has several immediate childnodes whose concepts belong to different hierarchies of the associated ontology, then we attribute the common ancestor of these concepts to node *n*.

After applying these propagation rules to the specific structure of document *d*, all nodes of this structure are associated with concepts of the selected ontology namely *Data Warehouse*. Thus, we obtain the semantic structure of the document.

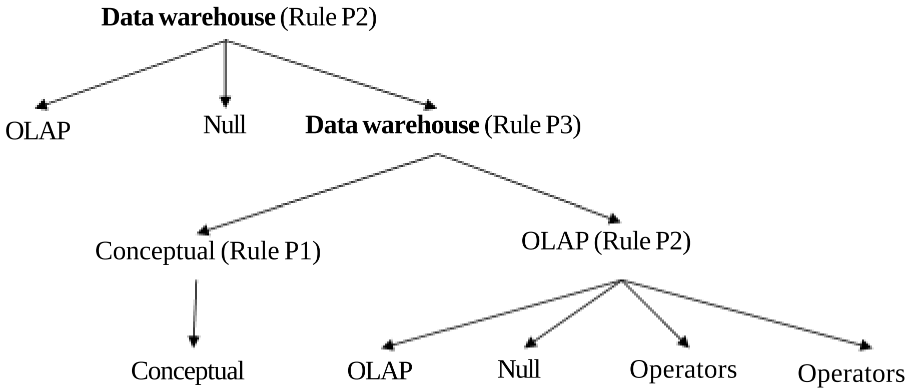


Figure 9. Semantic structure after propagation of concepts (in bold)

3.4 Pruning of semantic structure

This pruning phase is to keep the elements representing metadata, remove elements without assigned concept, etc. For this phase, we have defined the following set of four rules:

- **Rule Pr1:** Null elements, having no assigned concepts but corresponding to metadata of Dublin Core (such as author) will be maintained in the semantic structure.

- **Rule Pr2:** Null elements, having no assigned concepts and which are not metadata will be removed from the semantic structure.

- **Rule Pr3:** The successive elements assigned to the same concept will be replaced by a single element with the corresponding concept.

- **Rule Pr4:** Elements of a sub-tree that are assigned to the same concept will be replaced by a single element with the corresponding concept.

Changes operated to the structure of Figure 9 by applying these rules, are depicted in Figure 10.

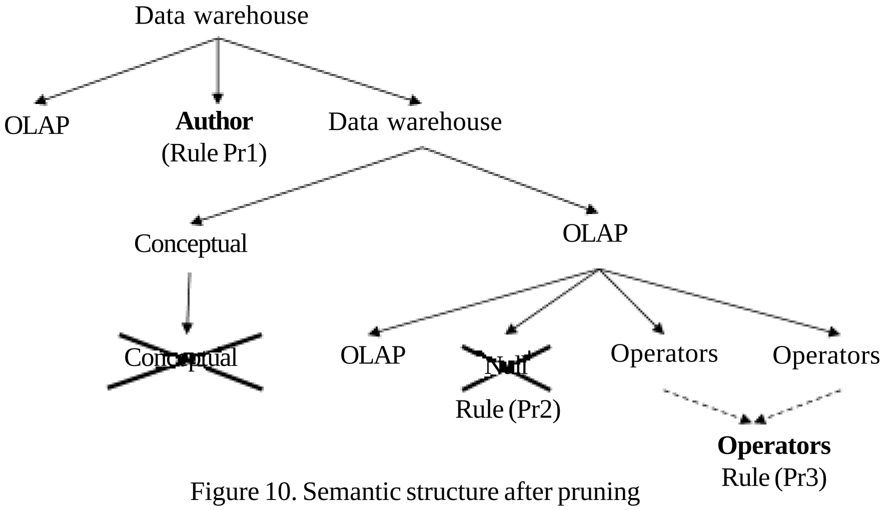


Figure 10. Semantic structure after pruning

Thus, the final semantic structure is shown in Figure 11.

4. Implementation

In order to validate the approach proposed in this paper, we are developing a software prototype. More accurately, we have

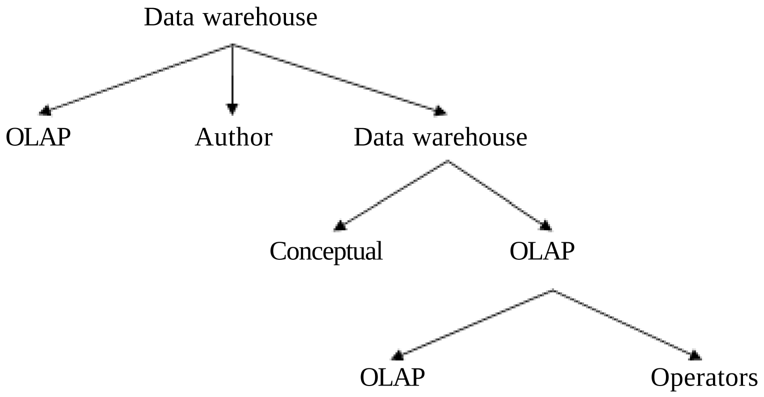
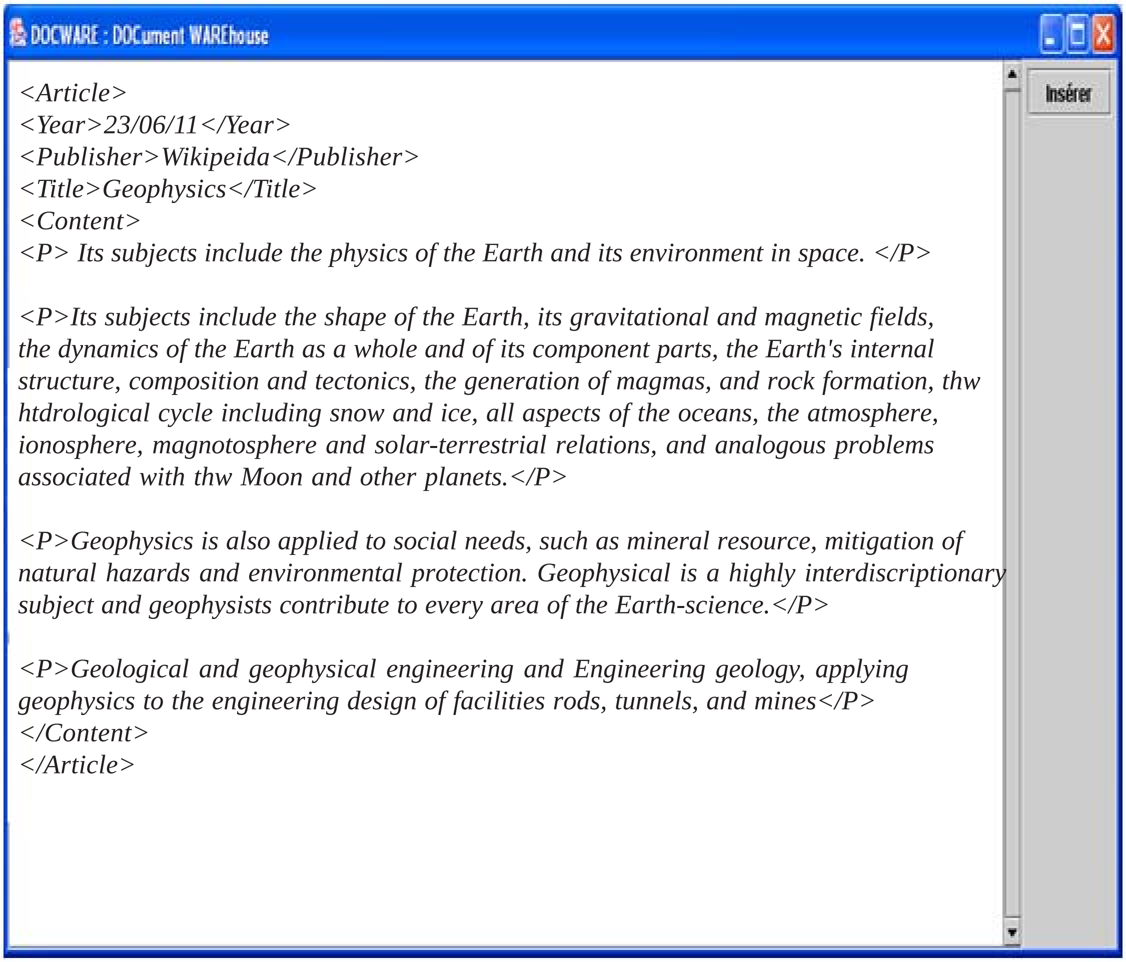


Figure 11. Semantic structure of the document *d*



```
<Article>
<Year>23/06/11</Year>
<Publisher>Wikipeida</Publisher>
<Title>Geophysics</Title>
<Content>
<P> Its subjects include the physics of the Earth and its environment in space. </P>

<P>Its subjects include the shape of the Earth, its gravitational and magnetic fields,
the dynamics of the Earth as a whole and of its component parts, the Earth's internal
structure, composition and tectonics, the generation of magmas, and rock formation, thw
htdrological cycle including snow and ice, all aspects of the oceans, the atmosphere,
ionosphere, magnotosphere and solar-terrestrial relations, and analogous problems
associated with thw Moon and other planets.</P>

<P>Geophysics is also applied to social needs, such as mineral resource, mitigation of
natural hazards and environmental protection. Geophysical is a highly interdiscriptionary
subject and geophysists contribute to every area of the Earth-science.</P>

<P>Geological and geophysical engineering and Engineering geology, applying
geophysics to the engineering design of facilities rods, tunnels, and mines</P>
</Content>
</Article>
```

Figure 12. A sample XML document

finished programming the first four phases. For that, we used:

- **JDOM:** FZor the extraction and manipulation of the logical structure of XML documents,
- **Java Parser:** for the implementation of the various steps for determining semantic structures,

- **Oracle DBMS:** for storing documents (content, logical and semantic structures).

This prototype enables to visualize for each document: its initial content (cf. Figure 12), its logical structure (cf. Figure 13) and its semantic structure (cf. Figure 14).

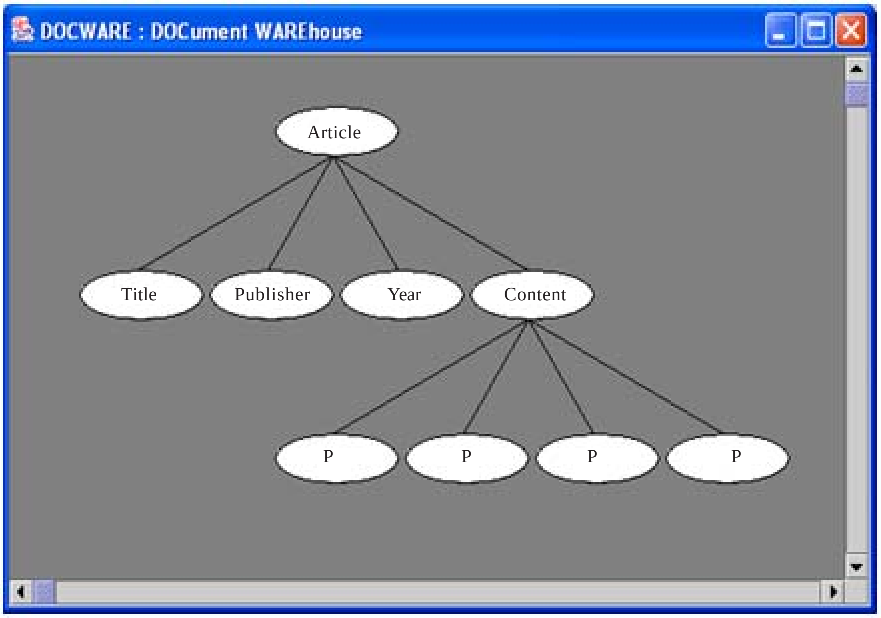


Figure 13. Logical structure of the XML document in Figure 12

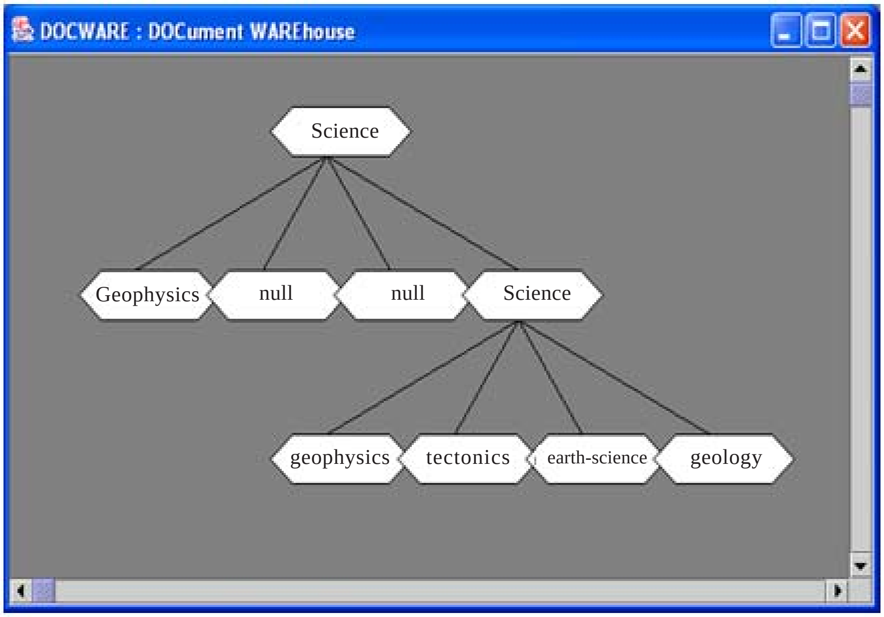


Figure 14. Semantic structure of the XML document in Figure 12

5. Conclusion

In this work, we have presented an automatic approach for the definition of the semantic structure for an XML document relying on the logical structure and the content of the document.

Our approach applies five steps. It starts by extracting the significant terms from leaf elements (textual fragments) of an XML document: this builds a logical structure for the document. Then, it determines which ontology matches better with the document in order to elaborate its semantic structure. This matching is realized through a specific metric. After that, it associates to each element of the logical structure of the document, a significant concept of the assigned ontology; this assignment is based on a calculation of the degree of similarity between the textual content and the concepts of the selected ontology. Finally, a pruning phase is envisaged, it cleans the semantic structure by removing noisy nodes (Null elements and successive elements assigned to the same concept) through a set of four pruning rules.

Several perspectives to this work constitute further challenges. One important issue consists of multi-semantic modeling of documents; this is because a single document generally interests several users, each of which may have a different perception/ comprehension of the document contents according to his/her domain. In addition, the integration of personalization aspects in the proposed approach is a promising idea. It is in order to improve query results by selecting for the user which semantic structures better interest him/her.

References

[1] Baeza-Yates, R., Ribero-Neto, B. (1999). Modern Information Retrieval, Addison Wesley.

[2] Baziz, M., Boughanem, M., Aussenac-Gilles, N. (2005). A Conceptual Indexing Approach based on Document Content Representation, *International Conference on Conceptions of Libraries and Information Science* (CoLIS), Glasgow, UK.

[3] Berisha-Bohé, S., Rumpler, B., Abascal, R. (2005). A Semantic Structure to Improve Information Retrieval Using XML, Conference on Electronic Publishing (ELPUB 2005), University of Leuven, Belgium.

[4] Ben Messaoud, I., Feki, I., Khrouf, K., Zurfluh, G. (2011). Unification of XML Document structures for Document Warehouse (DocW). *International Conference on Entreprise Information Systems* (ICEIS'11), 8-11 June, p. 85-94, Beijing, China.

[5] Gesche, S., Egyed-Zsigmond, E., Calabretto, S., Caplat, Beney, J. (2010). Classification supervisée sémantique d'articles de presse en français, Atelier Recherche d'Information Sémantique, Marseille, France.

[6] Harrathi, R., Calabretto, S. (2010). Une approche de recherche sémantique dans les documents semi-structurés, Atelier Recherche d'Information Sémantique, Marseille, France.

[7] Khrouf, K., Feki, J., Soulé-Dupuy, C. (2011). An Approach for Multidimensional Analysis of Documents, *International Conference on Information Systems and Economic Intelligence*, p. 46-53, Marrakech, Maroc.

[8] Salton, G., Fox, G. H., Wu, H. (1983). Introduction to Modern Information Retrieval, McGraw Hill International Book Company.

[9] Tagarelli, A., Greco, S. (2010). Semantic clustering of XML documents, *ACM Transactions on Information Systems* (TOIS), 28 (1) January.

[10] Upasana, P., Chakraverty, S., Rahul, J. (2010). Context Driven Technique for Document Classification, *International Conference on Advances in Computer Science*, India.

[11] Zargayouna, H., Salotti, S. (2004). SemIndex: a model of semantic indexing on XML documents, *European Conference on Information Retrieval* (ECIR'2004), Sunderland, UK, Avril.