International Journal of Electrical and Computer Engineering (IJECE)
Vol. 5, No. 5, October 2015, pp. 1174~1179
ISSN: 2088-8708

XML and Semantics

Mohammad Moradi*, Mohammad Reza Keyvanpour**
* Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran
** Department of Computer Engineering, Alzahra University, Tehran, Iran

Article Info

Article history:

Received May 8, 2015
Revised Jul 5, 2015
Accepted Jul 26, 2015

Keyword:

Metadata
Semantic annotation
Semantic Web
XML
XML semantics

ABSTRACT

Since the early days of introducing eXtensible Markup Language (XML), owing to its expressive capabilities and flexibilities, it became the defacto standard for representing, storing, and interchanging data on the Web. Such features have made XML one of the building blocks of the Semantic Web. From another viewpoint, since XML documents could be considered from content, structural, and semantic aspects, leveraging their semantics is very useful and applicable in different domains. However, XML does not by itself introduce any built-in mechanisms for governing semantics. For this reason, many studies have been conducted on the representation of semantics within/from XML documents. This paper studies and discusses different aspects of the mentioned topic in the form of an overview with an emphasis on the state of semantics in XML and its presentation methods.

Corresponding Author:

Mohammad Moradi,
Faculty of Computer and Information Technology Engineering,
Qazvin Branch, Islamic Azad University,
Qazvin Islamic Azad University - nokhbegan Blvd. Qazvin, Iran,
Email: Mhd.moradi@qiau.ac.ir

1. INTRODUCTION

Since the early days of invention and owing to its applicable features, XML [1] has become very popular in various applications in different domains. In fact, because of its flexibility and extensibility to different domains, it is known as defacto standard of publishing, storing, and exchanging data among (heterogeneous) systems and platforms, specifically the Web. Such interesting features have made XML one of the building blocks of the Semantic Web [2] and shortly incorporated it into different applications including semantic data integration, modeling, and creation of markup/description languages.

On the other hand, although XML allows users (programmers) to define their own tags and structures, it does not introduce any intrinsic and standard mechanism for representing semantics [3]. In this regard and due to invaluable applications of XML semantics, researchers have proposed several methods to leverage such semantics with respect to different features and capabilities of XML.

In this paper, the mentioned topic is studied and discussed from different aspects with a special emphasis on the state of semantics in XML and its applications and presentation methods. The structure of the paper is as follows: Section 2 provides a brief introduction of XML. In section 3, position of XML in Semantic Web is discussed. In section 4, the relation of XML and semantics and are studied. The notion of XML semantics and its different aspects and types are discussed in section 5. Sections 6 and 7 discuss some important considerations towards XML semantics and future works, respectively.

2. A GLIMPSE ON XML

XML as a subset of standard generalized markup language (SGML) was developed by the World Wide Web Consortium (W3C) in 1996 to become a cross platform, human and machine readable, and easy-to-create and -publish standard [1]. Despite HTML that is suitable for representing data (and information), XML directly deals with data in a non-presentational manner, specifically storing and interchanging data. Probably, one of the most interesting and applicable features of XML is the possibility of creating custom tags by users (document creators). Such a characteristic highly increases flexibility of XML documents, since users are not limited to a set of predefined (and probably meaningless tags) for building up their documents. Thus, at least in contrast with HTML documents, XML ones are more readable by humans and machines. Nonetheless, XML provides means for defining grammars and structures and does not introduce predefined rules and mechanisms to represent and interpret semantics within documents.

3. ROLE OF XML IN SEMANTIC WEB

Bringing semantics into the Web to shape the Semantic Web [2] was a milestone in its life. In fact, adding semantics to the content (and possibly structure) takes it to a new level of understandability by both humans and machines through the incorporation of meaning into content. Further, Semantic Web enables machines to comprehend semantic documents and data [2].In this regard, different data-centeric applications may leverage such semantics, such as data mining [4] and recommender systems [5].

One of the important issues about this field is the formalization of represented meanings (semantics) in order for standardization, domain-wide acceptability, and increasing portability and reusability. As a solution, Ontologies (documents or files that formally define the relations among terms [2]) have been proposed (and mostly utilized in the form of annotation) for different domains and application areas.

As mentioned earlier, XML is one of the major technologies for building Semantic Web which aims to provide an easy-to-use syntax for web data. Although it introduces the capabilities of encoding all kinds of data that are exchanged among systems [6], there are no mechanisms to interpret data or represent meanings in a structured and formal way.

In this way, besides the syntax and grammar (presented by XML), resource description framework (RDF) as a complementary technology for expressing the meanings was introduced [7]. RDF provides a standard, formalized, and interoperable model to describe facts about web resources, which gives some interpretations to the data [6].

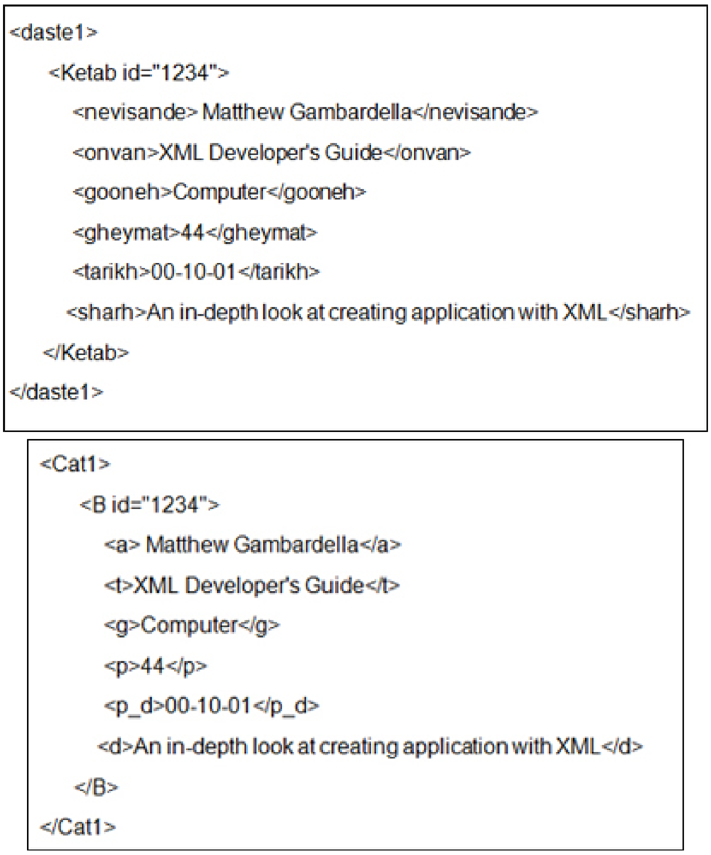
To explain the role of XML and RDF and their relationships in Semantic Web, it could be simply said that XML is responsible for syntax (data interchange), while RDF is thought to be a metadata data model.

4. XML AND SEMANTICS

Due to the lack of a widely-accepted and general standard, XML’s flexibility allows users to create their own tags freely and without any predefined limitation in contrast to HTML. Although these tags are sometimes meaningful, there is no guarantee that they could be intelligible by those who have no knowledge about the domain, or non-creators of documents. Figure 1 illustrates some examples of such XML documents.

Hence, compared with Semantic Web principles, XML cannot be introduced as a semantic markup language in its current form, but provides functionality to add meanings to the content in an unstructured way. In other words, the ideal is to create documents that are readable (and recognizable) by both human users and machines according to the intention of their creators. Nonetheless, this point is not nowadays perfectly realized.

Since using arbitrary tags is allowed in XML, in most cases, some types of implicit (but not formalized) semantics may be found within the documents. In fact, this is the reason that XML in public, known as a semantic markup language. Nonetheless, there are many application areas that take benefits of semantic aspect of XML including XML mining [8, 9] and XML retrieval [10, 11].



```
<daste1>
    <Ketab id="1234">
       <nevisande> Matthew Gambardella</nevisande>
       <onvan>XML Developer's Guide</onvan>
       <gooneh>Computer</gooneh>
       <gheymat>44</gheymat>
       <tarikh>00-10-01</tarikh>
      <sharh>An in-depth look at creating application with XML</sharh>
   </Ketab>
</daste1>
```

```
<Cat1>
    <B id="1234">
       <a> Matthew Gambardella</a>
       <t>XML Developer's Guide</t>
       <g>Computer</g>
       <p>44</p>
       <p_d>00-10-01</p_d>
      <d>An in-depth look at creating application with XML</d>
   </B>
</Cat1>
```

Figure 1. Examples of XML documents with Fingilish (Persian words written in Latin Alphabets) and vague tags

5. XML SEMANTICS

There are several issues about the quality of XML semantics; even some know XML only as a markup language- rather than a semantic one- that provides facilities to improve the semantics through markup via proposing the possibility of introducing user-defined tags [12]. Since XML has no predefined application-level processing semantics, it formally governs only syntax but not semantics [13].

Nonetheless, different levels of semantics could be seen in XML documents that may be leveraged in different application domains; thus, the main task of dealing with XML semantics is how to control it.

Due to the controversiality of the topic, to cope with the intrinsic issues, in the recent decade, researchers have performed studies in different directions. The important goal of such studies have been to propose methods for representing the semantics of XML documents in a formal and usable way to avoid ambiguity and capture implicit semantics within the documents.

According to the literature, the major directions of the researches around this topic could be understood as follows (Figure 2):

- Mapping (transforming) XML documents into RDF or OWL to represent their semantics (as in [14-17])
- Adding formal and organized semantics to the XML documents via semantic annotation or additional attributes and structures (as in [18-22])
- Extracting implicit semantics within the XML documents (as in [23-26])

These three main approaches are compared in Table 1.

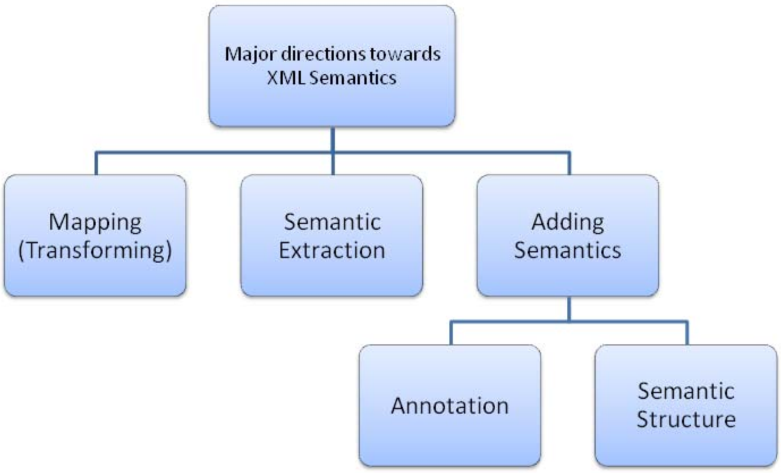


Figure 2. Classification of major directions around XML semantics

Although the common problem of the mentioned approaches is lack of a widely-accepted standard, it can be proved that methods that belong to the class of adding semantics to the XML documents have more advantages than others. In fact, mapping to and extraction of semantics have several essential issues that make them less-effective and, in some cases, impractical.

For example, while dealing with vague, ill-formed, and invalid documents, such approaches face serious problems. These approaches are usually used for dealing with legacy documents or when there is no control on marking-up documents (post-markup approaches). While adding semantics is feasible and applicable for producing documents (pre-markup approach).

Nevertheless, such techniques may be used to add semantics to the existing documents in some cases. In the latter approach, semantic annotation has several advantages over adding semantic structures to documents, which include preserving the natural structure of documents, producing lesser additional markup (pieces of text), and being easily expandable with minimum side effects on documents.

Table 1. Comparison of major directions around XML semantics

Approach	Benefits	Drawbacks	Implementation	Efficiency	Challenges
Mapping (Transforming)	Taking advantages of available semantic resources and facilities	Incomplete mapping, Dealing with invalid and vague documents	Relative (depends on the case)	Medium	Precise mapping, Choosing appropriate Ontologies, Semantic matching
Adding Semantics	Presenting controlled semantics, High level of precision	Producing additional markup, error prone	Easiest	Higher	Choosing appropriate (efficient) technique, Semantic comprehensiveness, standardization
Extraction of Semantics	Independency from external resources	Unguaranteed results, low level of precision	Hardest	Lower	Dealing with complex (nested), ambiguous and ill-formed documents, dealing with multilingual documents

6. CONSIDERATIONS

Owing to the intrinsic features of XML, it could be said that, almost in every document, there are some types of semantics, whether implicit or explicit. In this regard, taking a closer look at different levels of semantics within XML documents is a prerequisite for further actions to leverage them. From a general perspective, XML semantics may appear at three levels:

1. Element
2. Document
3. Group of documents

An XML element (or generically tag) may include attributes, text, or other elements. Accordingly, element's tag and/or its attribute(s) could be employed to express some type of semantics to present meaningful tags/attributes. This type of semantic (meaning) expression is the most straight forward and simplest. In most of XML documents, there are meaningful elements/attributes that are created deliberately or intentionally; but, the problem is that such instances are not usually formalized and, thus, are not recognizable and interpretable by machines.

Document-level semantics refers to the fact that, in some cases, some type of meaning may be inferred from the documents when they are considered as a whole. In other words, analyzing all elements of a given document could reveal some facts about its theme or purpose that may be regarded as semantics, specifically structural semantics.

Extending previously-mentioned concepts among several related (and most likely homogeneous) documents may present more meaningful information both on their content and structure. Such related documents usually share similar schema.

7. FUTURE WORKS

Based on the mentioned issues and approaches toward different aspects of semantics in XML documents, it is revealed that, despite its applications in a broad range of domains there is a vast gap between potential and actual capabilities and opportunity of XML in the representation of semantics. In fact, currently, the presentational (syntactical) aspect of XML is mostly used to freely express the content and structure for different purposes.

In this regard, as the future work, we will propose a new semantic annotation method for XML documents by leveraging intrinsic features of XML, namely attributes. Then, we take benefits of such sort of semantics for mining XML documents.

8. CONCLUSION

Owing to the high degree of flexibility and extensibility introduced by XML, shortly after proposing, it has become a popular means to store, represent, and interchange data between systems. Such expressive power also makes XML one of the building blocks of Semantic Web.

On the other hand, in contrast to the public thought, XML by itself does not present any mechanisms for governing semantics. Thus, there are several works on different aspects of this topic which provide solutions for dealing with XML semantics. In this paper, these issues and the current state of semantics in XML documents have been studied in the form of a concise overview. This work could be considered as a current state report of the topic for further studies.

REFERENCES

[1] T. Bray, *et al.*, “Extensible markup language (XML)”, [online] *World Wide Web Consortium Recommendation REC-xml-19980210*, www. w3. org/TR/1998/REC-xml-19980210 (1998, Accessed: 18 May 2014).

[2] T. Berners-Lee, *et al.*, “The Semantic Web”, *Scientific American*, vol. 284, no. 5, pp. 28-37, 2001.

[3] T. Kudrass, “Coping with semantics in XML document management”, in *Proceedings of the Ninth OOPSLA Workshop on Behavioral Semantics, Northeastern University*, 2001, pp. 150-161.

[4] F.B. Foroutan and H. Khotanlou, “Improving semantic clustering using with Ontology and rules”, *International Journal of Electrical and Computer Engineering (IJECE)*, vol.4, no. 1, pp. 7-15, 2014.

[5] K.B. Fard, *et al.*, “Recommender system based on semantic similarity”, *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 3, no. 6, pp. 751-761, 2013.

[6] M. Klein, “XML, RDF, and relatives”, *IEEE Intelligent Systems*, vol.16, no. 2, pp. 26-28, 2001.

[7] O. Lassila and R.Swick, “ Resource Description Framework (RDF) Model and Syntax Specification”, [online] *W3C Recommendation*,1999; http://www.w3.org/TR/REC-rdf-syntax/ (Accessed: 24 June 2014).

[8] A. Tagarelli and S. Greco, “Semantic clustering of XML documents”, *ACM Transactions on Information Systems (TOIS)*, vol. 28, no. 1, article 3, 2010.

[9] J.W. Lee, *et al.*, “Preparations for semantics-based XML mining”, in *Proceddings of IEEE International Conference on Data Mining, ICDM*, 2001, pp. 345-352.

[10] Q. Wang, *et al.*, “Exploiting semantic tags in XML retrieval”, in *Focused Retrieval and Evaluation, S.Geva, et al., Eds. Berlin: Springer Berlin Heidelberg*, 2010, pp. 133-144.

[11] D. Buscaldi, *et al.*, “Tag semantics for the retrieval of XML documents”, in *Proceedings of the 1st international symposium on Information and communication techNologies (ISICT '03), Dublin, Ireland*, 2003, pp. 273-278.

[12] R. Cover, “XML and Semantic Transparency”, *Technology Reports*, 1998, Available at: http://xml.coverpages.org/xmlAndSemantics.html, (Accessed 8 October 2014).

[13] H. Bohring and S.Auer, “Mapping XML to OWL Ontologies”, *Leipziger Informatik-Tage*, vol.72, pp. 147-156, 2005.

[14] V. Gancheva, "XML to RDF Scientific Data Transformation", in *Proceedings of the 5th European Computing Conference (ECC'11), Paris, France*, 2011, pp. 354-357.

[15] D. Van Deursen, *et al.*, "XML to RDF conversion: a generic approach", in *Proceedings of International Conference on Automated solutions for Cross Media Content and Multi-channel Distribution, AXMEDIS'08, Florence, Italia*, 2008, pp. 138-144.

[16] T. Rodrigues, *et al.*, "Mapping XML to Exiting OWL ontologies", in *Proceedings of International Conference WWW/Internet*, 2006, pp. 72-77.

[17] S. Liu, *et al.*, "XSDL: Making xml semantics explicit", in *Semantic Web and Databases, C. Bussler, et al., Eds. Berlin: Springer Berlin Heidelberg*, 2005, pp. 64-83.

[18] Y. Chen, *et al.*, " Expression of XML Implicit Semantics", in *Proceedings of The Second International Symposium on Networking and Network Security (ISNNS 2010), Jinggangshan, China*, 2010, pp. 15-19.

[19] G. Hignette, *et al.*, "Fuzzy semantic anNotation of xml documents", in *Proceedings of the CAiSE'05 WORKSHOPS, The 17th conference on advanced information systems engineering, DisWeb'05,Porto, Portugal*, 2005, pp. 319-332.

[20] F. Goasdoué, *et al.*, "Growing triples on trees: an XML-RDF hybrid model for annotated documents", *The VLDB Journal*, vol.22, no.5, pp. 589 -613, 2013.

[21] Y. Kotb, *et al.*, "XML Semantics", in *A. Scime, Ed, Web Mining: Applications and Techniques. Hershey, PA: Idea Group Publishing*, 2005, pp. 169-188.

[22] A. Renear, *et al.*, "Towards a semantics for XML markup", in *Proceedings of the 2002 ACM symposium on Document engineering, DocEng '02, McLean, VA, USA*, 2002, pp. 119-126.

[23] N. Aussenac-Gilles and M. Kamel, "Ontology Learning by Analyzing XML Document Structure and Content," in *Proceedings of the International Joint Conference on KNowledge Discovery, KNowledge Engineering and KNowledge Management, KEOD, Madeira, Portugal*, 2009, pp. 159-165.

[24] Y.Q. Yang, *et al.*, "An automatic semantic extraction algorithm for XML document", in *Proceedings of International Conference on Machine Vision and Human- Machine Interface (MVHI), Kaifeng, China*, 2010, pp. 41-44.

[25] L. Li, *et al.*, "Discovering semantics from data-centric XML", in *H. Decker, et al., Eds, Database and Expert Systems Applications. Springer Berlin Heidelberg*, 2013, pp. 88-102.

[26] S. Yang, *et al.*, "Derivation of OWL Ontology from XML Documents by Formal Semantic Modeling", *Journal of Computers*, vol. 8, no. 2, 2013.

BIOGRAPHIES OF AUTHORS

Mohammad Moradi received his B.S. in Software Engineering from Ghazali Higher education Institute, Qazvin, Iran. Currently, he is pursuing M.S. in Software Engineering at Islamic Azad University, Qazvin Branch, Qazvin, Iran. He is interested in Semantic Web and web 2.0 related topics as well as social networks and data mining.

Mohammad Reza Keyvanpour is an Assistant Professor at Alzahra University, Tehran, Iran. He received his B.S. in Software Engineering from Iran University of Science &Technology, Tehran, Iran. He received his M.S. and Ph.D. in Software Engineering from Tarbiat Modares University, Tehran, Iran. His research interests include image retrieval and data mining.