Comparative Analysis of XML Schema Languages for Improved Entropy Metric

*Kehinde Sotonwa
Department of Computer Science and Information Technology, Bells University of Technology, Nigeria
kehindeadasotonwa@gmail.com

Submitted: 12-SEP-2019; Reviewed: 22-OCT-2019; Accepted: 09-NOV-2019

Abstract- The eXtensible Markup Language (XML) is a data set to represent data in a format that is both human readable and machine readable. For XML documents to provide understanding about data exchange between applications, XML schema documents should be validated against the schema language. Most existing schema metrics were implemented differently in Document Type Definition (DTD), XML Schema Definition (XSD) and Regular Language for Next Generation (RNG) but never compare XML schema languages on any metric. Hence this paper compared three different schema languages on Improved Entropy Metric (IEM) using the Number of Attributes (NA), Number of Equivalence Class (FOC) and Number of Elements (NE). The proposed metric was applied on real schemas documents data are acquired from Web Service Description Language (WSDL) and implemented in DTD, XSD and RNG. The result showed that RNG reduce complexity of class elements, reflect strong support for class elements to appear in any order which showed more reusability and flexibility traits and overall understanding of the schema documents becomes much easier because RNG can be algorithmically converted and partner with other schema languages therefore this reduces maintenance effort.

Keywords- XML Schema Language, Schema Documents, Schema Metrics.

1 INTRODUCTION

The eXtensible Markup Language is playing important role in the exchange of variety of data on the web. XML as a new technology for web application to distribute data over the internet requires well design XML schema which is formed as tree structure contains root (parent element) and branches (child/sub-child elements). In developing web application some qualities such as reusable, flexible and maintainable must be considered to determine the complexity of application software in terms of entropy (Cerami, 2002; Erl, 2004; Newcomer and Lomow, 2004; Thomas, 2004).

The quantity of information contained in a document is evaluated as entropy which refers to disorder or uncertainty in a data set (Ruelia, 2012; Pathria and Beale, 2011; Shannon, 1948). Due to this Improved Entropy Metric (IEM) is formulated to see the schema language that can reduce complexity and still retain reusable, flexible and maintainable features. To ensure proper data exchange between applications XML documents must be validated against the XML schema language.

XML schema language is a description of a type of XML document, typically expressed in terms of constraints on the structure and content of documents of that type, above and beyond the basic syntactical constraints imposed by XML itself (Makoto et al., 2001). These constraints are generally expressed using some combination of grammatical rules governing the order of elements. Though there are a number of schema languages available, the three primary schema languages are DTD (Bray, Jean and Sperberg-McQueen, 2004); XSD (Binstock et al. 2002; Thompson et al. 2004) and RNG (Makoto 2000, ISO/IEC, 2002). Each language has its own advantages and disadvantages (Marconi and Nentwich, 2004).

2 TYPES OF SCHEMA LANGUAGES

2.1 DTD SCHEMA LANGUAGE

A DTD is a set of markup declarations that define a document type (DOCTYPE) for an SGML-family. It is the document structure with a list of legal elements and attributes (Steven, 2002). DTDs are the only one that can actually be embedded directly into the document. It can define data elements that can be used in the document; It does not support the namespaces, it supports only the text string data type, it is not object oriented hence, the concept of inheritance cannot be applied and it is also limited to express the cardinality for elements (Clark, 2003).

2.2 XSD SCHEMA LANGUAGE

This was the first separate schema language for XML to achieve recommendation status by the W3C. It can be used to verify each piece of item content in a document (Bikakis et al., 2014; Gao and Sperberg-McQueen, 2012; Thompson et al., 2004). It facilitates the design of open and extensible vocabularies to meet the requirement of data-oriented applications for a richer data typing system. XSD provide much greater specificity than DTDs could. However, the fit with the tree structure of XML documents is sometimes difficult to make. It is generally considered partly complex and hard to learn, it can be quite verbose (Guthuta, 2005; Clark, 2003). XSD does not implement most of the DTD ability to provide data elements to a document (Michael, 2009).

2.3 RNG SCHEMA LANGUAGE

RNG is an easy-to-learn schema language that possesses both XML syntax and compact non-XMl syntax (Clark and Makoto, 2001). This language can specify patterns for the structure and content of an XML document in a relatively simple but powerful way. It allows attributes to be treated as elements in content models (Dongwon and Wesley, 2000). Most RNG schemas can be algorithmically converted into XSD and even DTDs. Nevertheless, RNG has no ability to apply default attribute data to an element's list of attributes (Makoto et al., 2005).

*Corresponding Author
Table 1 show the comparison in term of features between the three (3) schema languages: XSD language, DTD language and RNG language.

Table 1. Comparison of the Schema Languages (DTD XSD and RNG)

<table>
<thead>
<tr>
<th>Language Features</th>
<th>DTD</th>
<th>XSD</th>
<th>RNG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Namespaces Support</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Multiple top level elements</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Ordered contents</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Simple datatypes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Schema Modularity</td>
<td>Some</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Content model extensibility</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Content dependency</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Pattern matching</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Any namespaces (nesting)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Sibling content</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Mutual exclusion</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

3 RELATED WORKS

Klettke, Scneider and Heuer (2002) proposed some well-known metrics like LOC, McCabe, Fan-in and Fan-out, Depth of Inheritance Tree (DIT), adapted for measuring the complexity of DTD. The metric failed to evaluate if a DTD is data centric or document centric. McDowell, Schmidt and Yue (2004) proposed eleven metrics for XSD and two formulae used these metrics to calculate the quality indices for XSD and the complexity indices to conform XML documents. The future work is to measure the number of app Info elements, which provide documentation about the XML schemas to applications. Visser 2006 also adopted some well-known existing metrics developed for other software artifacts to XSD to deal with the programs and grammar of structural complexity of XSD.

Basci and Misra (2010 and 2011), proposed metrics which followed a similar approach taken by Davis and LeBlanc (1988) metric for the assessment of the structural complexity based on schema entropy concept and intended to measure the interface complexity of the schema documents in XSD and DTD but the metric failed to reflect the reusability of the schema documents in comparing schemas of equal number of complex type definitions and to address the issue of limited possibilities of expressing class element in any order which have different sizes.

Thaw and Khin (2013) measured the reusable, extensible and understandable qualities of XML schema documents in XSD. The measures were formulated based on binary entropy function and rank order centroid method. The drawback of these measures is that no software tool has been implemented to aid in their measurement. Thaw and Misra (2013) formulated Entropy Measure of Complexity metric (EMC) intended to measure the reusable quality of XML schema documents based on the entropy concept, inheritance feature elements and attributes. But EMC did not satisfy the additive properties thus making the metric complex and was not adopted in the industry.

Falola et al., (2017) evaluated and made comparison of metrics for XML schema language which is based on their unique features, advantages and limitations. In addition, the study also discussed whether or not theoretical, practical and empirical validations had been conducted on the various metrics that were implemented in DTD and XSD but not RNG. Sotonwa et al., (2019) measured SLOC metric for RNG schema documents to predict the amount of effort required to develop a program and give details of all line of codes; whether helping in the efficiency of code execution or not. Sotonwa et al., (2019) developed improved metrics to measures schema entropy and interface complexity implemented in RNG but did not compare other schema languages.

4 RESEARCH METHODOLOGY

The concept of entropy has been used in literature to express decline in quality, reliability, maintainability and understandability of software due to diversity in the structures of elements. The Schema Entropy metric proposed by (Basci and Misra, 2010) is defined as:

\[
SE = - \sum_{i=1}^{n} (P(C_i) \log_2(P(C_i))
\]

where \(n\) is the number of distinct classes and \(P(C_i)\) is the probability of class

\[
IEM = - \sum_{i=1}^{NEC} (FOC_i \log_2(FOC_i)) + NA
\]

Where

- NEC: Number of Equivalence Class reflects the number of unique element structures in the schema documents.
- FOC: Frequency of Occurrence is the member count of each class which reflects the number of occurrences of each class member.
- NA: Number of Attribute is the number of features used to describe a property or to provide additional information about an element in the schema documents.

4.1 VALIDATECARD IN DTD, XSD AND RNG

ValidateCard is a link acquired from the WSDL used as sample of the demonstration of the proposed metric; implemented in DTD, XSD and RNG with different elements such as: validate card, card number, type, response and result that show the position of each of the element and attribute if any; belonging to in the tree structure of the directed graphs representation as shown in Fig. 1, 3 and 5 and their Listings as the interpretation of the graph in different class number e.g. C1-C6 seen in Fig. 2, 4 and 6 while the analyses are given blow:
4.2 ANALYSIS OF IEM

i. $IEM = DTD_{\text{ValidateCard}}$

$$IEM = \sum_{i} (FOC_i \log_2 (FOC_i) + NA)$$

$$= - \left[ \left( \frac{1}{6} \right) \ast \log_2 \left( \frac{1}{6} \right) + \left( \frac{1}{6} \right) \ast \log_2 \left( \frac{1}{6} \right) + \left( \frac{3}{6} \right) \ast \log_2 \left( \frac{3}{6} \right) \right] + 0$$

$$= 1.7920$$

ii. $IEM = XSD_{\text{ValidateCard}}$

$$IEM = \sum_{i} (FOC_i \log_2 (FOC_i) + NA)$$

$$= - \left[ \left( \frac{1}{6} \right) \ast \log_2 \left( \frac{1}{6} \right) + \left( \frac{1}{6} \right) \ast \log_2 \left( \frac{1}{6} \right) + \left( \frac{2}{6} \right) \ast \log_2 \left( \frac{2}{6} \right) + \left( \frac{2}{6} \right) \ast \log_2 \left( \frac{2}{6} \right) \right] + 0$$

$$= 1.9179$$

iii. $IEM = RNG_{\text{ValidateCard}}$

$$IEM = \sum_{i} (FOC_i \log_2 (FOC_i) + NA)$$

$$= - \left[ \left( \frac{1}{6} \right) \ast \log_2 \left( \frac{1}{6} \right) + \left( \frac{2}{6} \right) \ast \log_2 \left( \frac{2}{6} \right) + \left( \frac{3}{6} \right) \ast \log_2 \left( \frac{3}{6} \right) \right] + 0$$

$$= 1.4589$$

5 RESULT AND DISCUSSION

Series of experiments were conducted to show the effectiveness of schema language using IEM, as performance measurement. Analyses of all the implemented DTD, XSD and RNG can be seen in Table 2.

Table 2. Complexity Measure for DTD, XSD and RNG

<table>
<thead>
<tr>
<th>Schema Document/No</th>
<th>NA</th>
<th>DTD</th>
<th>XSD</th>
<th>RNG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1.7920</td>
<td>1.9179</td>
<td>1.4589</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2.9920</td>
<td>2.4349</td>
<td>2.4349</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>2.1275</td>
<td>2.0715</td>
<td>1.4845</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>2.5221</td>
<td>2.5221</td>
<td>1.7220</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1.9220</td>
<td>1.5000</td>
<td>1.8549</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>2.2510</td>
<td>2.2512</td>
<td>2.2512</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>5.0469</td>
<td>4.8806</td>
<td>4.2605</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1.6640</td>
<td>1.8113</td>
<td>1.4056</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>2.2170</td>
<td>2.7769</td>
<td>1.4956</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>2.0001</td>
<td>0.1812</td>
<td>0.0182</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>4.9717</td>
<td>4.6750</td>
<td>3.2513</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>4.5947</td>
<td>4.5947</td>
<td>3.2863</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>4.5803</td>
<td>2.7642</td>
<td>2.0782</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>2.3221</td>
<td>0.7219</td>
<td>0.7219</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>6.9468</td>
<td>6.7292</td>
<td>5.4690</td>
</tr>
</tbody>
</table>

5.1 COMPARATIVE STUDY OF DTD, XSD AND RNG SCHEMA LANGUAGES

The relative graphs in Fig. 7 shows the comparison of the IEM of DTD, XSD and RNG. Close inspections of this graph shows different complexities values for DTD, XSD and RNG. RNG gives lower complexity values for all schema documents because it provides strong support for unordered content by allowing sequence of pattern to appear in any order; because high FOC makes the developer more familiar to schema documents due to reusability feature therefore overall understandability of schema documents becomes much easier and this require less maintenance effort while XSD and DTD languages do not allow sequence of pattern to appear in any order.
Fig. 8 shows the comparison between DTD and RNG, since DTD is not object oriented, the concept of inheritance cannot be applied, thus give high complexity values of IEM of DTD while RNG is an object oriented and the concept of inheritance is applied therefore, IEM can capture decreasing physiological complexity of RNG due to familiarity gained by navigating the child and sub child many times.

The graph in Fig. 9 showed comparison for RNG and XSD. Both allow for similar mechanisms of specificity, both allow a degree of modularity in their languages for splitting the schema into multiple files and both of them are, or can be, defined in an XML language thus, they have their complexities values closer for all schema documents and especially similar complexity values for schema documents 6 and 10 (ConvetTemp and Bank) as 2.2512 and 0.9182. Lastly, Fig. 10 showed comparison between DTD and XSD schema languages. XSD support the features of namespace awareness, responsiveness and data types to provide alternative to DTD but DTD does not support these and this gives reason to greater complexity values to DTD than XSD in some schema documents.

6 CONCLUSION

This paper work proposed Improved Entropy Metric (IEM) based on information contained in the WSDL of the schema documents. The proposed metric considered fundamental factors which directly affect the complexity of the schema document, it used frequency occurrence of the class (FOC), number of element (NE), number of equivalence class (NEC), number of attributes (NA) and i<sup>th</sup> class, since information is contained in the elements and attributes of the schema documents. The study also presented different schema languages measure IEM techniques which were analyzed and the limitations of the schema languages were discussed.

The IEM make more sensitive measurement in understanding the information content contained in the schema documents. The applicability of the metric was evaluated by different schemas implemented in DTD, XSD and RND to prove its robustness and effectiveness. The difficulty in understanding the schema documents was measured and the results showed that RNG is a more suitable language when compared with DTD and XSD.

DTD and XSD were not able to measure class elements comprehension; of a fact, empirical validation has shown that RNG is able to reflect strong support for class elements to make them appear in any order. XSD and DTD cannot be algorithmically converted to other schema language; RNG language permits such. Lastly, RNG is highly structured and can partner with other schema...
language with a separate data typing language which makes it simpler in exhibiting a better presentation of a given schema document than DTD and XSD.

ACKNOWLEDGMENTS

I acknowledge the God the father, the Son and the Holy Spirit for making this work possible. To God be thy Glory. My sincere appreciation goes to entire staff of the Department of Computer Science and Engineering LAUTECH, Ogbomoso, Nigeria and Department of Computer Science and Information Technology, BELLSTECH, Ota, Nigeria for their encouragement and supports.

REFERENCES


Michael S. (2009): DEVX article Taking XML Validation to the Next Level: Introducing CAM


Steven Holzner. (2002): Sams Teach Yourself XML in 10 minutes. Sams Publishing, 00 East 96th Street, Indianapolis, Indiana, 46240 USA.


**APPENDIX**

**DTD Code for ValidateCard**

```xml
<xml encoding="UTF-8">
<ELEMENT ValidateCard(ValidateCardNumber,ValidateCardNumberResponse)>
  <ELEMENT ValidateCardNumber (cardType,cardNumber)*>
    <ELEMENT ValidateCardNumberResponse (ValidateCardNumberResult)*>
    <ELEMENT cardType (#PCDATA)>
    <ELEMENT cardNumber (#PCDATA)>
    <ELEMENT ValidateCardNumberResult (#PCDATA)>
</ELEMENT>
</DOCUMENT>
```

**XSD Codes for ValidateCardNumber**

```xml
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  <xs:element name="ValidateCard">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="ValidateCardNumber"/>
        <xs:element ref="ValidateCardNumberResponse"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="ValidateCardNumber">
    <xs:complexType>
      <xs:sequence minOccurs="0" maxOccurs="unbounded">
        <xs:element ref="cardType"/>
        <xs:element ref="cardNumber"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="ValidateCardNumberResponse">
    <xs:complexType>
      <xs:sequence>
        <xs:element minOccurs="0" maxOccurs="unbounded" ref="ValidateCardNumberResult"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="ValidateCardNumberResult" type="xs:string"/>
</xs:schema>
```

**RNG Codes for ValidateCard**

```xml
<?xml version="1.0" encoding="UTF-8"?>
<grammar xmlns="http://relaxng.org/ns/structure/1.0"
  xmlns:a="http://relaxng.org/ns/compatibility/annotations/1.0"
  datatypelibrary="http://www.w3.org/2001/XMLSchema-datatypes">
  <start>
    <element name="ValidateCard">
      <element name="ValidateCardNumber">
        <zeroOrMore>
          <element name="cardType" type="xs:string"/>
        </zeroOrMore>
        <element name="ValidateCardNumberResult" type="xs:string"/>
    </element>
</grammar>
```

**LINKS OF WSDL FOR SCHEMA DOCUMENTS**

- http://webservices.daellab.net/temperatureconversions/TemperatureConversions.wso?WSDL
- http://www.esendex.co.uk/secure/messenger/soap/InboxService.asmxml?WSDL
- http://services.argosoft.com/AddressValidation/AddressVerifier.asmx?WSDL
- http://hooch.cis.gsu.edu/bgates/MathStuff/Mathservice.asmx?WSDL