Entropy Measure of Quality Metrics for XML Schema Documents

Tin Zar Thaw
University of Computer Study, Mandalay
tinzar.t@gmail.com

Abstract

Extensible Markup Language (XML) Schema documents (XSD) play an important role in software development process and need to be qualified with software qualities. A good quality design of XSD increases software productivity and minimize development time. In this respect, there are many qualities: maintainability, reusability, understandability, extensibility and so on. Among these qualities, this paper focuses on two qualities: extensible and reusable qualities and proposes two metrics to measure these qualities of schema documents based on Entropy method. Moreover, due to the defined maximum and minimum values for each quality metric, the proposed metrics can measure the specified quality. Therefore, the metrics can provide valuable information for improving the quality of XML based systems. The usefulness of the present metrics is empirically proved through actual test cases.

Keywords- *Software Development Process, XML Schema, Entropy Method, Extensibility, and Reusability*

1. Introduction and Related Works

Many different domains, organizations and content providers have been publishing and exchanging information via internet by the usage of eXtensible Markup Language (XML) and standard schemas. Design of XSD plays an extremely important role in software development process and needs to be quantified for many software qualities such as reusability, expendability, understandability, maintainability and so on. Many metrics have been developed with respect to improving quality in the software engineering process. Metrics are useful in software development process and help developer to ensure their product with particular qualities. The design and specification of XML schemas should be as rigorous an activity as designing and developing code or designing database schemas. As such, when creating an XML schema we should be working within a development process and working to a set of design guidelines and coding standards. XML schemas should be reviewed for accuracy and compliance with guidelines and standards [6].

Research of metrics for XSD is scarce. There has been considerable research done with respect to improving quality in the software engineering process and discovering best practices for knowledge and data capture. A. McDowell, C. Schmidt and K. Bun Yue [1] proposed and discussed eleven metrics to measure the quality and complexity of XML Schema and conforming XML documents. When developing the metrics, they also focused on the categories of the ISO 9126 quality model. An open source metric analyzer tool for XML Schema had been developed. Their tool was easily been extended to add new metrics and altered the composition of the indices to best fit the requirements of a given application. An important future task was the validation or refutation of the current formulae to determine the relative complexity and quality of XML Schema documents.

One research that has been done was by D. Basci and S. Misra [2]. In it, they developed a new complexity metric, (XSD), to measure of the complexity of XML schema documents and their proposed metric was based on the internal architecture of the XSD components and hence considered the complexities of its building components. They demonstrated the validity of their complexity metrics theoretically. These researchers have being done complexity metric and shown the validity of their complexity metric theoretically and empirically using xml schema documents. For the validation of the proposed metric they analyzed 65 real example Schema files from the web. If the C(XSD) value increased, it was clear that the schema quality was decreased since it implied inefficient use of memory and time [4].

Another research has been done by the same researchers. In their paper, they formulated a metric for the assessment of the structural complexity of XSD. Their proposed metric ‘Schema Entropy’ was based on entropy concept and intended to measure the complexity of the schema documents written in W3C XML Schema Language due to diversity in the structures of its elements. Their metric provided valuable information about the reliability and maintainability of systems. In this respect, their metric was believed to be a valuable contribution for improving the quality of XML-based systems. It was demonstrated with Examples (projects.xsd, books.xsd and others) and validated empirically through actual test cases [3].

2. Prosed Metrics: Reusable Quality (RQ) and Extensible Quality (EQ)

The proposed metrics are based on the binary entropy function and used to measure extensible and reusable qualities. A set of design guidelines provides information about these two qualities. Nowadays, the longer lived and more widely used XML schema is going to be, the more important it is to include extensibility in XML schema. Extensible XML schemas should be designed to easily allow for additions at a later date [8]. But, increasing extensibility tends to increasing complexity of XML schema documents. Moreover, Reusable XML schemas should be specified in such a way that types and elements can be leveraged by other XML schemas [6]. Therefore, every type defined in an XML schema that is the content type of an attribute or an element should be defined globally. Types that are defined globally can be reused in other XML schemas. The proposed metrics are explained with 2 parts. Firstly, part A presents collection of the based attributes to calculate the proposed metrics. Part B shows the formulation of the proposed metrics and identification of the maximum and minimum values for each quality.

2.1. Collection of Based Metrics

The proposed quality metrics are adaptation of existing metrics for software qualities. Before calculating based metrics, we firstly built the document object model (DOM) trees for schema documents. Over the DOM trees, the based metrics are counted. All based metrics are defined as the following and its value will be assumed as 0.1 if each metric is zero. Then we assume that all qualities’ values of all based metrics directly depend on their total number of elements for a given schema except the IISp and Mr metrics. IISp metric value depends on a fourth of the total number of elements in a given document and the Mr metric gets full quality if this value is equal to and greater than two.

(i) Total Number of Substitution Groups (SG)

Substitution groups provide a mechanism for XML elements that is similar to subtype polymorphism in programming languages [5].

It simply provides a mechanism for allowing elements to be used interchangeable. Substitution groups allow schema authors to create or utilize schemas that define generic base types and extend these types to be more domain-specific without affecting the original schema.

(ii) Total Number of Union and Any (U&A)

Placing an <xs:any> at the end of the complex type content is a good way of adding extensibility to the XML schema. Union types enables an element or attribute value to be one or more instances of one type drawn from the union of multiple atomic and list types [6]. For extensible quality view, U&A metric is calculated as:

U&Q = Total number of any components + Total number of union components.

(iii) Total Number of Import, Include and Namespace (IISp)

Extension schemas are XML schemas that extend one namespace by providing (in another namespace) types, elements, and attributes designed to provide extra features and work with this schema. Hiding namespaces moves the complexity of a document's framework to the Schema level. Additionally, maintenance is easier when hiding namespaces as it is possible to change a Schema without impact to instance documents [5]. As a result, the IISp metric of a given schema document is formulated as the below.

IISp = Total number of Include components + Total number of Import components + Total number of Name Spaces

(iv) Number of global elements (Eg)

This metric measures the total number of global elements in the given schema document. For maximal reusability and extendibility, elements should be defined globally [6].

(v) Average number of user defined type references (Mr)

For the reusable purpose, a user defined type should be created for so many elements and attributes in XSD [5]. The Mr metric is calculated as:

Mr = Total Number of Type Declaration / Total Number of Type Definition

(vi) Number of reuse type (Tr)

The element types can be used for the reusable purpose. But, if complex types or simple types are anonymous, they can’t be the target for derivation and can’t be used by later versions of the schema. So, this metric is formulated as:

Tr = Number of user-defined type declarations - Number of anonymous types

(vii) Number of inheritance types (Ti)

In object oriented programming, inheritance is a way to compartmentalize and reuse code by creating collection of attributes and elements. In XML schema languages, types are defined with extension and restriction key words. This metric is calculated as:

Ti = Number of simple types by restriction + number of complex types by restriction + number of complex types by extension

(viii) Total number of elements (Te) and total number of user defined type declaration (T)

These two metrics are needed to obtain reasonable quality values for the whole given document.

2.2. Formulation of Proposed Metrics and Identification of Maximum and Minimum Values

The proposed metrics are formulated based on Binary Entropy Function that characterizes the purity of the collection of design structure components for schema document [7]. Given a collection S, containing positive and negative examples of some target concept, the entropy of S relative to this Boolean classification is

$$\text{Entropy(S)} \equiv - \; p_{\oplus} \log_2 p_{\oplus} + p_{\ominus} \log_2 p_{\ominus} \; \ldots (1)$$

where $p_{\oplus}$ is the proportion of positive examples in S and p_{Θ} is the proportion of negative examples in S. In all calculation involving entropy they define $0\log_2 0$ to be 0. The entropy is 0 if all members of S belong to the same class. Note that the entropy is 1 when the collection contains an equal number of positive and negative examples. If the collection contains unequal numbers of positive and negative examples, the entropy is between 0 and 1.

According to guidelines, we assume that all based metrics have the same priority for the certain target. Generally, in this paper, to measure two qualities we prepare the equation (1). This prepared formula is:

$$\text{Entropy(TQ)=} - \sum_{i=1}^{n} (p_{i\oplus} \log_2 p_{i\oplus} + p_{i\ominus} \log_2 p_{i\ominus}) \ldots (2)$$

Where TQ is the target quality: extendible quality (EQ) and reusable quality (RQ). $p_{\oplus}$ is a half proportion of positive feature components of a class and p_{Θ} is a proportion of the sum of negative feature components and a half positive feature components of this class. The Entropy of the target quality having i distinct classes where i= 1, 2,…n and n is the number of based attributes for a target quality. The EQ and RQ metric contain 4 based attributes. The classes of the EQ metric are IISp, Ti, U&A and SG metrics. The RQ metric contains Eg, Mr, Tr and Ti metrics and they are very useful for this quality. Each class contains a collection of the same components that has positive or negative features. In order to measure the percentages of each target quality for schema documents, we need to identify the maximum and minimum values for each quality that the maximum value of the target quality is equal to the total number of classes and the minimum value is equal to zero for all qualities. Identifying these values the proposed metrics can have desirable mathematical properties.

(1) A metric value can be in a meaningful range (e.g., minimum value to maximum value, where minimum value truly means quality absence, maximum value indicates that the input schema is in full quality, and 0.5 represents the "half-way point") .

(2) A metric represents a characteristic that increases when positive traits occur or decrease when undesirable traits are encountered.

For each class, Entropy is 0 when a component collection contains all negative feature components. Then increasing positive feature components tend to increasing Entropy value. Finally, the highest value of an Entropy class is 1 when this collection contains all positive feature components.

3. Empirical Result

Empirical validation proves the practical utility of the proposed metric. To check the validity of the proposed metrics: the dc.xsd document is downloaded from the link:

http://www.dublincore.org/schemas/xmls/qdc/2008/02/11/dc.xsd. The graph representation of the Schema documents that have more similarly-structured elements with higher frequencies of occurrences exhibit more regularity, thus are easy to grasp because of gained familiarity. Firstly, the EQ metric exploits a directed graph representation of a schema document, known as the G(SD). The directed graph representation of the Schema document, G(SD), can be defined as G(SD) = (N,E), where N is a set of nodes representing the elements of XSD and can be defined as: *N = <N1, N2,..., Ni>, i = 1,2...n* where *n* is the total number of components (attribute and element) in SD; E is a set of edges that represent parent-child relationships between the components of XSD. These components that have no child component are represented by leaf nodes and connected to their associated component nodes having circle shapes by straight lines in G(SD). Circle nodes represent elements and one rectangle shows one group of the given documents. The directed graph representation of this document is shown in figure 1.

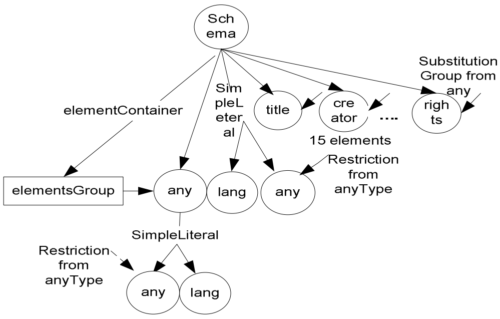

Figure 1. The directed graph of dc.xsd

In this figure, the text on the edge is type name of node and the text of the tail of the arrow represent substitution group, restriction and extension from other component. The dc.xsd document contains 21 element nodes. For example, the RQ and EQ metrics values for the schema document *dc.xsd* are calculated by using Equation (2):

Eg ⟵ [16,7], Mr ⟵ [1,20], Tr ⟵ [2,19] and Ti ⟵ [2,19]

Entropy (RQ) = 1.6734289184484032

IISp ⟵ [2,19], Ti ⟵ [4,17], U&A ⟵ [19,2] and SG ⟵ [15,6]

Entropy (EQ) =2.663644963885717

To check the validity of the proposed metric, the 60 files are downloaded from the well-known websites. Some of the analyzed Schemas were extracted from the Web Service Description Language and the other files contained mathematical nature information. These files are analyzed and calculated the EQ and RQ values according to Equation (2). To analyze the accuracy of two metrics, we have calculated the average accuracies of varying file numbers such as 5, 10, 15,…, 60. The results are shown in Figure.1 and Figure.2. According to their results, we can see that the accuracy of RQ metric is increasing when the number of files is increasing. The accuracy value is over 80% for all changing the total number of files. The accuracy of EQ metric fluctuates with increasing towards the maximum number of files. But, overall testing, this accuracy value is always over 80%. The RQ metric is suitable for measuring the reusable quality. For extensibility, the EQ metric can also produce acceptable result because of the overall accuracy. But, the system needs to be tested with many files until the accuracy value is stable. Therefore, Binary Entropy Function is useful for measuring Reusable Quality and Extensible Quality on the XML schema documents.

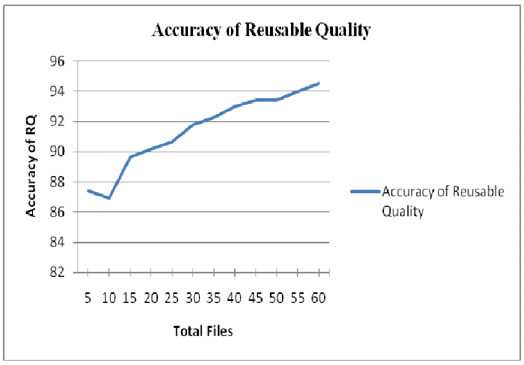

Figure 2. The accuracy result of the reusable quality metric.

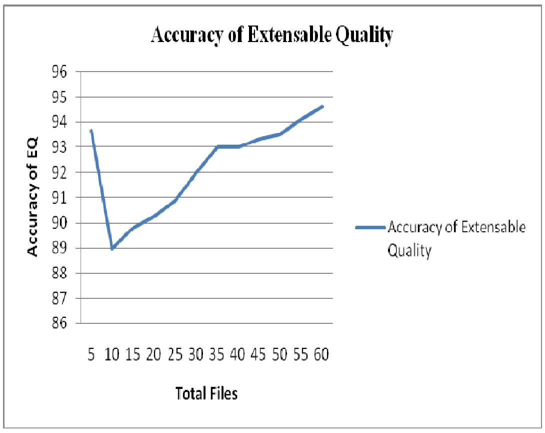

Figure 3. The accuracy result of the extensible quality metric.

5. CONCLUSIONS

In this paper, we have presented eight metrics for XML Schema and proposed two quality metrics based on them and Binary Entropy function. These two metric are also qualifying with two properties. In order to check the reliability of quality metrics, 49 files are downloaded and evaluated empirically using 4 test cases. It is found that the EQ and RQ can predict Extensible and Reusable Qualities with more than 80 percent accuracy for schema documents respectively. Moreover, the proposed metrics can provide valuable information for improving the quality of XML based system.

References

[1] D. Basci and S. Misra, "Complexity Metric for XML Schema Documents" in Proceedings of the ExampleGraph25th International Workshop on SOA and Web Practices, 2007, pp. 1-14.

[2] D. Basci and S. Misra, "Entropy as a Measure of Quality of XML Schema Document",*the International Arab Journal of Information Technology, Vol. 8, No. 1, January 2010.*

[3] D. Basci and S. Misra, "Measuring and Evaluating a Design Complexity Metric for XML Schema Documents" , Journal of Information Science and Engineering 25, 1405-1425 (2009).

[4] T. M. Cover and J. A. Thomas, "Elements of Information Theory", Second Edition (2006).

[5] A. McDowell, C. Schmidt and K. Bun Yue, "Analysis and Metrics of XML Schema", In Proceedings of Intl Conference on Software Engineering Research and Practice,pp. 538-544 (2004).

[6] D. Obasanjo, "W3C XML Schema Desing Patterns: Avoiding Complexity" , Microsoft Corporation, january 2004, Originally published on http://www.xml.com.

[7] D. Stephenson, "XML Schema Best Practice", December 2004 Hewlett-Packard Development Company.

[8] *A publication of the Postsecondary Electronic Standards Council (PESC), " PESC Guidelines for XML Architecture and Data Modeling", version 3.0, april 29, 2005.*