Journal of Software Engineering and Applications, 2013, 6, 458-469
http://dx.doi.org/10.4236/jsea.2013.69056 Published Online September 2013 (http://www.scirp.org/journal/jsea)

Measuring Qualities of XML Schema Documents

Tin Zar Thaw, Mie Mie Khin

Department of Computer Science, University of Computer Studies, Mandalay, Myanmar.
Email: tinzar.t@gmail.com

Received January 29th, 2013; revised March 1st, 2013; accepted March 9th, 2013

ABSTRACT

The Extensible Markup Language (XML) is becoming a de-facto standard for exchanging information among the web applications. Efficient implementation of web application needs to be efficient implementation of XML and XML schema document. The quality of XML document has great impact on the design quality of its schema document. Therefore, the design of XML schema document plays an important role in web engineering process and needs to have many schema qualities: functionality, extensibility, reusability, understandability, maintainability and so on. Three schema metrics: Reusable Quality metric (RQ), Extensible Quality metric (EQ) and Understandable Quality metric (UQ) are proposed to measure the Reusable, Extensible and Understandable of XML schema documents in web engineering process respectively. The base attributes are selected according to XML Quality Assurance Design Guidelines. These metrics are formulated based on Binary Entropy Function and Rank Order Centroid method. To check the validity of the proposed metrics empirically and analytically, the self-organizing feature map (SOM) and Weyuker's 9 properties are used.

Keywords: Extensible Markup Language; XML Schema Documents; Web Engineering Process; XML Quality Assurance Design Guidelines; Schema Qualities

1. Introduction

Nowadays, the Extensible Markup Language (XML) based web applications are widely used for data exchanging and network services. Efficient implementation of web application needs to be efficient implementation of XML and XML schema document. In web engineering process, metrics are used to measure the quality of the creating software but research of XML schema quality metrics is scare. The proposed metrics are formulated based on the Binary Entropy Function and rank order weight method.

XML Schema has performed two functions: specifying structure relationships of the internal components and specifying mechanisms for validating the content of XML component. Moreover, specifying relationships of XML schema components are very difficult for schema developers if the schema is becoming large. If new schema can be built from previously developed schemas and the schemas are easy to understand and extend for enhancing functions, correcting faults, or adapting them to new circumstances, schemas may save more time than poor quality schemas.

Information theory based software measures are attractive because they quantify the physiological complexity of a given Schema document [1]. Binary Entropy Function characterizes the purity of design structure components of a given schema document. The based attributes are identified as having more positive impact on particular qualities of XML schema. There are many rank order weighting methods: the equal weights (EW), rank sum (RS), rank reciprocal (RR), rank-order centroid (ROC) and rank methods. Among these methods, the rank-order centroid (ROC) is a well-known method and produces more accurate result than other methods. The proposed measures are evaluated empirically using the Kohonen Self-Organizing Maps (SOM) algorithm, a non-linear mapping, from a high-dimensional data space to a low-dimensional space. It can cluster efficiently without more understanding the meaning of input data [2].

The usefulness and quality of a new metric is evaluated by using validation process. Theoretical validation is the ensuring process of metrics with the principles of measurement theory and empirical evaluation is the study of software in order to characterize and predict. The two approaches complement one another; a valid set of metrics should be both theoretically validated and empirically evaluated. Therefore, a new metric must be evaluated formally and practically for its validation. In order to prove the validation of the presented metric, the metric is evaluated by Nine Weyuker's Properties [3]. The paper is organized as follows. Section 1 introduces about XML and its schema qualities. Section 2 presents the related researches about XML schema metrics. The attribute selection and formulation of proposed metrics are explained in Sections 3 and 4 respectively. The proposed metrics are proved empirically and theoretically in Sections 5 and 6. Section 7 concludes the paper.

2. Related Work

Nowadays, several metrics have been proposed for software developers and software development groups to measure the size and quality of software product during software development process. Many researchers proposed metrics for Document Type Description and Extensible Markup Language (XML) Schema.

The researchers proposed a metric to measure the complexity of Document Type Definition (DTD) documents based on the recursive relationship of elements: the number of elements, connectors, appearance indicators and back edges in a DTD. Their information assisted in comprehending the complexities presented in DTDs and DTD libraries in [4]. Other researchers presented two complexity metrics: Entropy metric and Distinct Structured Element Repetition Scale metric for measuring DTD documents. They compared two metrics with other metrics and they suggested that their metrics were useful in differentiating DTDs with the same size in [5]. Moreover, the most relevant research was done by [6]. In it, a set of five metrics developed for DTD documents to measure the complexity of XML documents and to concentrate on usability and maintainability. They were lines of code, McCabe complexity, structure depth, fan-in and fan-out metrics. They suggested that size and complexity can be applied to complete DTDs.

Bun Yue *et al.* [7] proposed two composite indices: quality index and complexity index of XML Schema documents focused on the ISO 9126 quality model. The paper [8] proposed a metric to measure the complexity due to the internal architecture and recursion of XML schema components. To validate the metric empirically by comparison with other metrics applied on XML schema documents. They suggested that the more memory and time uses efficiently, the better the schema quality is. The researchers in [9,10] proposed two metrics: Schema Entropy and the total complexity of the XML schema documents to measure the structural complexity of XML schema document. They suggested that their metric provided valuable information about the reliability and maintainability of systems. Visser *et al.* in [11] proposed a suite of metrics over graph representations of schema structure. These metrics were tree impurity, fan-in, fan-out and instability, efferent and afferent coupling and gain instability, coherence and normalized count of modules that were mostly adaptations of existing metrics for other software language. They suggested that the measurement results may be used for assessing potential risks in schemas. According to my knowledge, researchers have been done to determine complexity and maintainable quality of the XML schema documents. There is no the reusable, extensible and understandable qualities measuring. Therefore, in this paper, these quality metrics are proposed for software developers to facilitate private-assessment and improve of their schema based software products in terms of saving cost.

3. Attribute Selection for XML Qualities

Selecting the based attributes is great impact on measuring particular software qualities. In this paper, the base attributes are selected according to XML Quality Assurance Design Guidelines.

3.1. XML Qualities

There are many qualities for XML Schema documents: understandable, reusable, maintainable, extensible, flexible, compliable and so on [12-16]. Among these qualities, the reusable quality, extensible and understandable quailties are focus on this thesis.

- Reusable: An XM schema component can be globally defined and leveraged by other XML schemas and components in the same document. Reuse concept is that new schema should not be built from scratch, but should be able to leverage previously developed schemas.
- Extensible: Extensible quality xml schema is possible for developers to write an extension schema by adding additional features to the original in a controlled way.
- Understandable: If the system is difficult to understand for enhancing relationships between schema components or adapting them to new circumstances, changes may increase the development cost. XML schemas are clear, consistent and unambiguous having human readable components.

3.2. Attribute Selection

The best practices and the guidelines will allow XML Schema developers to develop good quality schemas. The based attributes are collected according to XML Quality Assurance Design Guidelines [3,4,12,14,16,17]. To get the reusable, extensible and understandable qualities of XML schema documents, many experts guide how attributes are used to get particular quality.

To achieve a high level of reusable quality, types and elements that are defined globally can be reused in other XML schemas and in the same XML schema documents. If complex types are anonymous, they can't provide reuse property. Inheritance has the property to compartmentalize and reuse the collections of schema elements and attributes by using keywords by extension are used to inheritance schema component structure. For the RQ metric, the following attributes are selected.

- Number of reuse types (Tr).

$$\begin{aligned}\text{Tr} = {} & \text{Number of user defined type declaration} \\ & - \text{Number of anonymous type declaration}\end{aligned} \tag{1}$$

- Number of inheritance types(Ti).

$$\begin{aligned}\text{Ti} = {} & \text{Number of simple type with restriction} \\ & + \text{Number of complex type with restriction} \\ & + \text{Number of complex type with extension}\end{aligned} \tag{2}$$

- Number of global elements (Eg).
- Average number of user defined type references (Mr).

$$\text{Mr} = \frac{\text{Number of type declaration}}{\text{Number of type definition}} \tag{3}$$

- Total number of types (Tt).
- Total number of elements (Te).

The inheritance feature type is also support for extensible quality of schema document. For extensible purpose, XML schema send and import types, elements, and attributes from one namespace into the main XML schema by using "include" and "import" tags respectively. The Union types also allow developers to combine types. A wildcard matches element and attribute information items dependent on their namespace name by using the <xs:any> tag to extend schema components. Substitution groups allow elements to be used interchangeable and support the extensible quality of schema document but having many substitution groups tends to more difficult process of these documents. For the extensible quality of XML schema documents, the following based attributes are selected.

- Number of include and import components (I&I).
- Number of inheritance types (Ti).
- Number of union and any (U&A).

$$\begin{aligned}U \,\&\, A = {} & \text{Number of union components} \\ & - \text{Number of any components}\end{aligned} \tag{4}$$

- Number of substitution groups(SG).
- Total Number of Types (Tt).
- Total Number of Elements (Te).

Understandable quality schema directly support for other software qualities: reusable, maintainability, extensible and so on. To get the understandable quality of schema documents, human readable schema components are very important for software developers and development group. XML Schema provides three understandable components: documentation, annotation and links to some information. Therefore, the UQ metric is proposed on the following based selected metrics.

- Number of annotations (Na).
- Number of documentation (Nd).
- Number of links to requirement and document (Nl).
- Total Number of Nodes (Tn).

3.3. Assigning Attributes' Weights

According to guidelines [12-14,16,18], the important of individual attributes are not the same, their weight is not equal. Moreover, many XML schema experts provide the judgments about the positive ranks of attributes for particular quality. Therefore, rank-order centroid (ROC) method is used to calculate weights of attributes for three metrics. In XML schema components, type definition is more important than element declaration. The selected attributes are ranked according to XML schema experts' judgments for each of quality.

For reusable quality, type definitions Tr and Ti have higher ranks than other attributes. Eg has the higher reuse than Mr. Tt is used to get the ratios of Tr and Ti and Te is used to get the ratios of Eg and Mr. For the RQ metrics, the selected attributes are ranked:

$$\text{Tr} \geq \text{Ti} \geq \text{Eg} \geq \text{Mr}$$

According to extensible purpose, I&I is more important than other internal components. Then type definition, Ti, is ranked. Having many substitution groups tends to more difficult schema process than having many unions. Tt and Te are used to get the ratios of above attributes. For the EQ metric, the selected attributes are ranked as follows:

$$\text{I\&I} \geq \text{Ti} \geq \text{SG} \geq \text{U\&A}$$

For the understandable quality, an annotation component is more important than documentation and link. Many schema documents are more used documentation component than link component. The Tn attribute is used to get the ratio of the ranked attributes. For the UQ metric, the selected attributes are ranked:

$$\text{Na} \geq \text{Nd} \geq \text{Nl}$$

Rank-Order Centroid (ROC) method:

$$ROC_i W(\mu) = \frac{1}{t}\sum_{l=i}^{t}\frac{1}{p_l},\ i = 1,2,\ldots,t \tag{5}$$

μ = a metric.
t = the total number of attributes for μ metric.
p_l = the position of i^{th} attribute for μ metric.

$$i \in \begin{cases} \{N_a, N_d, N_l\}\ \mu = UQ \\ \{I \,\&\, I, T_i, SG, U \,\&\, A\}\ \mu = EQ \\ \{T_r, T_i, E_g, M_r\}\ \mu = RQ \end{cases} \tag{6}$$

Example of Weight Calculation for the based attribute (Tr) of RQ metric:

$$ROC_{Tr}W\left(RQ\right)=\frac{1}{t}\sum_{l=i}^{t}\frac{1}{p_l},\ i=1,2,\dots,t$$

$$=\frac{1}{4}*\left(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}\right)=0.5208$$

4. Formulation of Proposed Metrics

The proposed three metrics are formulated based on the binary entropy Function and rank order centroid method to measure the qualities of XML schema documents. It contains two parts: formulation of proposed metrics and analyzing result of proposed metrics.

4.1. Formulation of Proposed Metrics

The proposed three metrics: RQ, EQ and UQ are formulated based on the binary entropy Function and rank order centroid method to measure qualities of a given schema document S. The binary entropy function of based attributes for a given single document is as follow [1]:

$$Entropy_{\,i}\left(S\right)=-p_{i^{\oplus}}\log_2 p_{i^{\oplus}}-p_{i^{\ominus}}\log_2 p_{i^{\ominus}} \tag{7}$$

where $p_{i\oplus}$ is the positive concept of i component or a half proportion of positive concept for i^{th} schema component and $p_{i\Theta}$ is the negative concept of i component. The positive and negative concepts are given below:

$$p_{i^{\oplus}}=\frac{i}{2D_i} \tag{8}$$

$$p_{i^{\ominus}}=\frac{D_i-\frac{i}{2}}{D_i} \tag{9}$$

where D_i is used to get the ratio concept of i schema component. The D_i is defined as follows:

$$D_i=\begin{cases} T_t, i=T_r \mid T_i \\ T_{n,}i=N_d \mid N_l \mid N_a \\ T_e,\left(\text{otherwise}\right) \end{cases} \tag{10}$$

Generally, to measure reusable, extensible and understandable qualities, the equation is formulated by multiplying with ROC weight method according to Equations (5) and (7):

$$\mu\left(S\right)=\left(\sum_{i=1}^{n}Entropy_i\left(S\right)\times ROC_iW\right)\times 100 \tag{11}$$

μ can be RQ, EQ or UQ metric. For each of metrics, the based attributes set according to Equation (6). In all calculation, the entropy value is 1 if the schema contains an equal number of positive and negative components. If the given schema contains unequal numbers of positive and negative components, the entropy is between 0 and 1. To demonstrate three metrics, the xml schema document of dc.xsd is analyzed and measured. This schema document is built as DOM tree and the based attributes are counted as follows.

{Tr→2, Ti→1, Eg→16, Mr→0.5}

Three qualities of XML schema document (dc.xsd) are measured according to Equation (6).

$$\begin{aligned} RQ_{i\in\{Tr,Ti,Eg,Mr\}} &= \left(\sum_{i=1}^{n}\left(Entropy_i\left(dc.xsd\right)\times\frac{1}{n}\sum_{l=i}^{n}\frac{1}{p_l}\right)\right)\times 100 \\ &= \left(\sum_{i=1}^{n}\left(-p_{i^{\oplus}}\log_2 p_{i^{\oplus}}-p_{i^{\ominus}}\log_2 p_{i^{\ominus}}\times\frac{1}{4}\sum_{l=i}^{n}\frac{1}{p_l}\right)\right)\times 100 \\ &= \left(-p_{Tr^{\oplus}}\log_2 p_{Tr^{\oplus}}-p_{Tr^{\ominus}}\log_2 p_{Tr^{\ominus}}\times\frac{1}{4}\sum_{l=1}^{4}\frac{1}{p_l}\right)\times 100 \\ &\quad+\left(-p_{Ti^{\oplus}}\log_2 p_{Ti^{\oplus}}-p_{Ti^{\ominus}}\log_2 p_{Ti^{\ominus}}\times\frac{1}{4}\sum_{l=2}^{4}\frac{1}{p_l}\right)\times 100 \\ &\quad+\left(-p_{Eg^{\oplus}}\log_2 p_{Eg^{\oplus}}-p_{Eg^{\ominus}}\log_2 p_{Eg^{\ominus}}\times\frac{1}{4}\sum_{l=3}^{4}\frac{1}{p_l}\right)\times 100 \\ &\quad+\left(-p_{Mr^{\oplus}}\log_2 p_{Mr^{\oplus}}-p_{Mr^{\ominus}}\log_2 p_{Mr^{\ominus}}\times\frac{1}{4}\sum_{l=4}^{4}\frac{1}{p_l}\right)\times 100 \end{aligned}$$

$$\begin{aligned} RQ_{i\in\{Tr,Ti,Eg,Mr\}} &= \left(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}\right)\times 0.5208\times 100+\left(-\frac{0.5}{2}\log_2\frac{0.5}{2}-\frac{1.5}{2}\log_2\frac{1.5}{2}\right)\times 0.2708\times 100 \\ &\quad+\left(-\frac{8}{19}\log_2\frac{8}{19}-\frac{11}{19}\log_2\frac{11}{19}\right)\times 0.1458\times 100+\left(-\frac{0.25}{19}\log_2\frac{0.25}{19}-\frac{18.75}{19}\log_2\frac{18.75}{19}\right)\times 0.0625\times 100 \end{aligned}$$

$$RQiW_{i\in\{Tr,Ti,Eg,Mr\}} = 88.9978$$

4.2. Analyzing Result of Proposed Metrics

The proposed metrics are analyzed by using 20 XML schema files and the binary entropy function is compared with simple ratio.

4.2.1. Comparison of Simple Ratio and Binary Entropy Function

Ratio and Binary Entropy Function Values are between 0 and 1. For example, a given schema document has the total components (20) and Number of Annotation (23). Total component of a give schema document is the sum of total elements, total types and a root component (schema node) of that document. Human understandable components: Number of Annotation, Number of Documentation and Number of links are used to explain types and elements of the same documents. It is clear that if one of human understandable components is larger than the total components for a given schema document, the result can be less than one. To solve this problem, in simple type ratio, the ratio is 23/20(1.15) and it is greater than one. Binary Entropy function can solve this problem as follows:

$$Entropy_{\,i}\left(S\right) = -p_{i\oplus}\log_2 p_{i\oplus} - p_{i\ominus}\log_2 p_{i\ominus}$$

$$\begin{aligned}&\text{Entropy}_{\,Na}\left(\text{ExampleSchema}\right)\\&= -p_{Na\oplus}\log_2 p_{Na\oplus} - p_{Na\oplus}\log_2 p_{Na\oplus} - p_{Na\ominus}\log_2 p_{Na\ominus}\\&= -\frac{11.5}{20}\log_2\frac{11.5}{20} - \frac{8.5}{20}\log_2\frac{8.5}{20}\\&= 0.983708\end{aligned}$$

The binary entropy function value is not exceeded one as well as can reduce the human readable value having more than total components. Therefore, the proposed metrics is formulated by summarization the binary entropy functions of based metrics and multiplying their rank weights.

4.2.2. Analyzing Result of Proposed Metrics

To analyze the proposed metrics, the attributes and three qualities of the following schema documents are analyzed and calculated. These schema document links are shown in **Table 1** in Appendix.

The correlation between reusable related attributes and the RQ metric value for reusable quality of XML schema documents are analyzed in **Figure 1**. There are the ratios of Tr and Ti to Tt and ratios of Eg and Mr to Te. In this analysis, the schemas: ID_34 has the highest value at 77.22 because the ratios of Tr and Ti (two third and one third of Tt), the ratio of Eg (half a Te), the ratio of Mr(23 times smaller than Te) . For the schema: ID_17, Te is the highest number but the Eg value is less than one third of Te and Mr is very low value. Moreover, Tr is eleven times smaller than Tt and Ti is seven times smaller. Therefore, the schema: ID_17 has lower reusable quality value than ID_34.

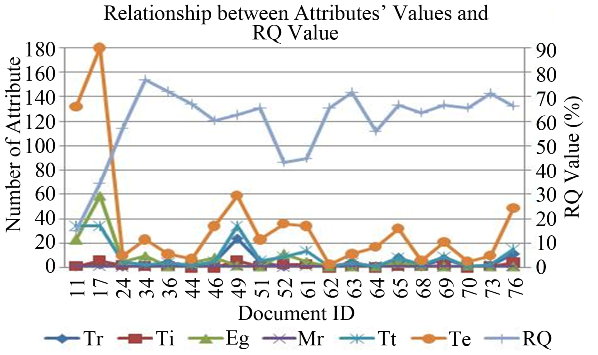


Figure 1. Reusable related attributes' values vs RQ metric value.

The correlation between extensible related attributes and the EQ metric value for extensible quality of XML schema documents are analyzed in **Figure 2**.

There are the I&I, U&A and SG ratios to Te and the Ti ratio to Tt. In this analysis, the XML schema document: ID_34 has higher value than other schema documents because of the overall highest ratio.

The correlation between human readable attributes and the UQ metric value for understandable quality of XML schema documents are analyzed in **Figure 3**. There are the ratios of Na, Nd and Nl to Tn. In this analysis, it is clear that higher human readable attributes' ratios tend to higher the understandable quality of XML schema document.

5. Empirical Validation of Proposed Metrics

To prove the validity and usefulness of the proposed metrics, the clustering method, Self-organizing maps (SOM) and well-known XML schemas are used. One hundred XML schemas are downloaded for validating the proposed metrics that are shown in **Table 1**. These schemas are: WSDL Schemas for describing network services, Mathematical Schemas for calculation, Metadata Vocabulary Schemas for monitoring and changing systems affecting everything in life and Geospatial Schemas for developers to ensure complex spatial information. XML schema documents are clustered by SOM depending on two types: only particular proposed metric and it's based attributes and the cluster matching results are compared. It is clear that the proposed metrics is validated and useful according to the SOM clustering result.

5.1. Self-Organizing Map Clustering

The unsupervised learning method, SOM is suitable for

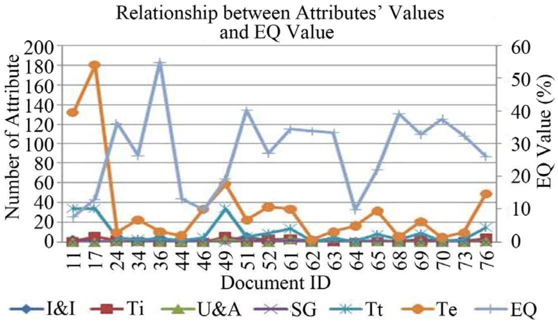


Figure 2. Reusable related attributes' values vs EQ metric value.

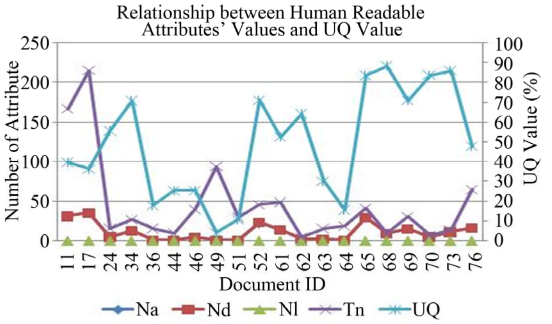


Figure 3. Human readable attributes' values vs UQ metric value.

measuring the validity and usefulness of the proposed metrics that they are difficult to prove their validity and usefulness because they don't have historical data and other comparison metrics. One advantage is that they do not require more understanding with input data. The SOM learns similarity of input feature vectors and responds groups of similar input vectors [2].

Before clustering XML schema documents, two data files are created for each of proposed metrics, one contains the particular proposed metric values of one hundred XML schema documents and another contains its based attributes of those. In preprocessing step, file attributes are converted into binary data according to Equations (12) and (13).

$$f(x)=\begin{cases}1, if\left(x>\dfrac{f(y)}{2}\right)\\ 0,(\text{otherwise})\end{cases} \tag{12}$$

where $f(x)$: the binary conversion function that converts the x based metric value into binary data and $f(y)$: the function that chooses T_t, T_e, T_n and $Max(T_n)$ based on y.

$$f(y)=\begin{cases}T_t, \text{else if } (y=tr\,|\,ti)\\ T_e, \text{else if } (y=Eg\,|\,Mr\,|\,U\,\&\,A\,|\,SG)\\ T_n, \text{else if } (y=Tt\,|\,Te\,|\,Na\,|\,Nd\,|\,Nl\,|\,I\,\&\,I)\\ MAX\,(T_n attributes)\,\text{otherwise}\end{cases} \tag{13}$$

After converting binary data, the contingency tables are created for each of six files. For creating the contingency table, there are the variables a, b, c and d, i^{th} position, the attribute x of a record and the attribute y of another record.

To create the contingency table, dissimilarities are calculated by Jaccard coefficient [6]:

$$d(x,y)=\frac{e}{e+f+g} \tag{14}$$

e = number of times $x_i = 1$ and $y_i = 1$.

f = number of times $x_i = 0$ and $y_i = 1$.

g = number of times $x_i = 1$ and $y_i = 0$.

These dissimilarity matrixes of schema files are trained with the SOM and the schema documents of each file are cluster in terms of dissimilarity matrixes. For instance, the two files of EMC metric are clustered by SOM and the cluster matching result is analyzed.

SOM clusters XML schema documents into two groups depending on the based attributes Na, Nd and Nl. The first group contains schema IDs: 65, 68, 70, 73, 96 and 99 and the second group contains other schema documents in **Figure 4**. Moreover, in **Figure 5**, SOM also clusters XML schema documents into two groups depending on only the UQ metric values. The first group contains schema IDs: 24, 34, 52, 61, 62, 65, 68, 69, 70 and 73 and the second group, other group contains other schema documents. The first groups' members of two

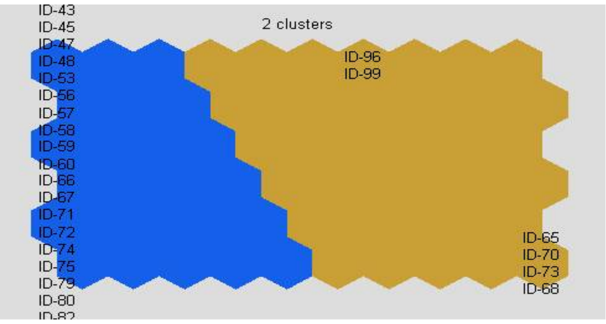


Figure 4. SOM clustering result depending on based attributes.

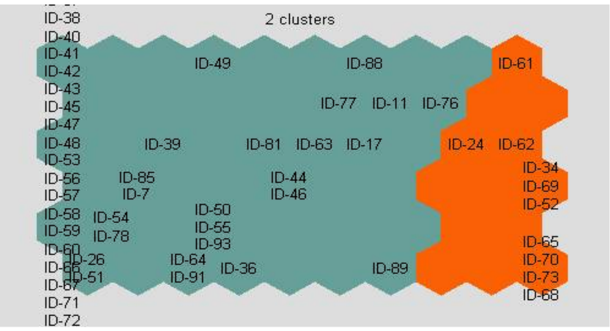


Figure 5. SOM clustering result depending on only proposed UQ metric.

types are then compared and analyzed. It is clear that four schema documents are matched and eight schema documents are not equal. Moreover, the second groups' members of two types are also compared and analyzed. The result shows that 82 schema documents are matched. Therefore the total cluster matching result is 82%.

For all proposed metrics, the cluster matching results are greater than 80 percent. Therefore, the proposed three metrics are useful to measure reusable, extensible and understandable qualities of XML schema documents.

5.2. Measuring and Evaluation of Proposed Metrics

To prove the validity and usefulness of the proposed RQ, EQ and UQ metrics, a hundred of standard XML schemas are used. To analyze the SOM cluster matching results, the cluster thresholds are assumed and are compared with SOM cluster thresholds of two types for each of four proposed metrics. The SOM clusters XML schema documents according to two types: depending on only the proposed metric and depending on the based attributes of the proposed metric and produces two groups for each of types. One group of a type is then compared with one group of other type and the cluster matching results of two types is produced.

According to **Figure 6**, 66 schema documents do not contained human readable components and therefore their understandable quality are zero. The SOM can produce these documents into one group in terms of only UQ value. But, the SOM cannot produce them into one group in term of based attributes and 2 documents contain in another group. It is clear that the SOM clustering result depending on only proposed UQ metric can produce more accurate result than depending on the based attributes of UQ metric. Moreover, the two SOM clustering results are compared with the assumed clustering result, the cluster matching results are 100% in terms of only UQ value and 92% in terms of the based metrics of UQ.

To analyze the clustering results of two types, 100 XML schema documents are assumed into two groups: from 7.8981 to 17.165 (13 schema documents) and from 17.165 to 77.124 (87 schema documents). The SOM clustering results of RQ metric are shown in **Figure 7**. Tr, Ti, Eg, Mr, Tt and Te are the based attribute of the RQ metric. The two SOM clustering results are compared with the assumed clustering result, the cluster matching results are 100% in terms of only RQ value and 85% in terms of the based metrics of RQ.

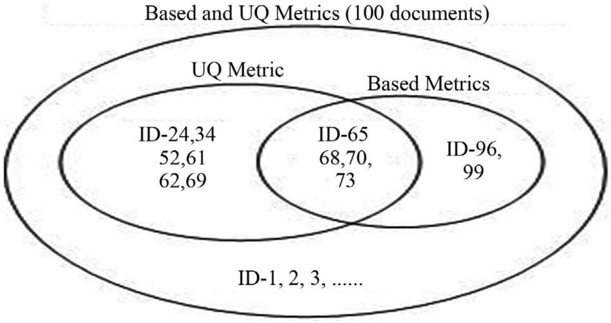


Figure 6. Cluster matching of UQ and its based metrics.

SOM clustering results of EQ metric are shown in **Figure 8**. I & I, Ti, SG, U & A, Tt and Te are the based attributes of the EQ metric. The two SOM clustering results are compared with the assumed clustering result, the cluster matching results are 100% in terms of only EQ value and 96% in terms of the based metrics of EQ.

According to the analyzed result empirically, the thee metrics are more suitable for measuring reusable, extensible and understandable qualities of XML schema documents than the based attributes because of matching result with more than 80%.

6. Analytical Validation of the Proposed Metrics

The usefulness and quality of a new metric is evaluated by using validation process analytically. To validate the proposed three metrics, the well known Weyuker's properties [11] are used.

Analytical Evaluation of the Proposed Metrics against Weyuker's Properties

In this section, the validation of the proposed metrics is

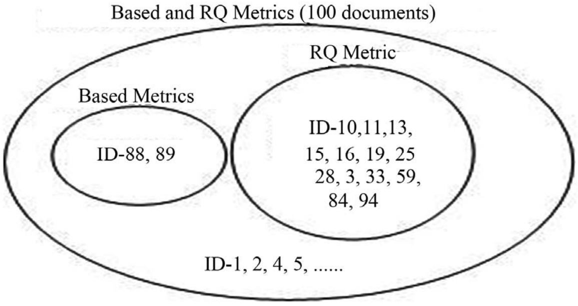


Figure 7. Cluster matching of RQ and its based metrics.

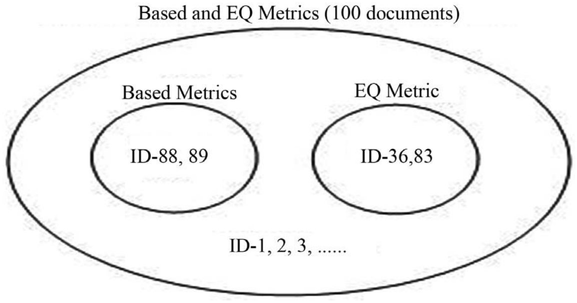


Figure 8. Cluster matching of EQ and its based metrics.

also proved by using nine well known Weyuker's properties through examples [12]. To validate the analytical validation, assume a metric (μ), three distinct schema documents (A, B and C) and the combined AB, AC and BC documents. For instance, the AB documents is combined the A document with the B document.

The evaluations of the proposed metric against Weyuker's properties are as follows.

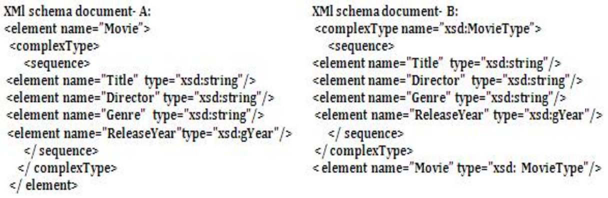
1) **Non-Coarseness (($\exists$ A) ($\exists$ B) (μ(A) ≠ μ(B))):** This means that a metric has not to measure and produce the same value for all schema documents. It is clear that three proposed metrics satisfy this property.

2) **Granularity: Let *c* be a non-negative number, and then there are only finite numbers of schema documents of complexity *c*.** This states that there are only a finite number of XML schema documents of the same complexity. The proposed metrics satisfy this property because of the same complexity values with the same ratios having the same rank.

3) **Non-uniqueness (μ(A) = μ(B)):** This means that there are distinct schema documents of the same complexity value. For example, in the **Table 1**, the schema A (ID_64) is metadata vocabulary schema and the schema B (ID_46) is a geospatial schema for developer to develop publicly available interface standards for spatial information. The EQ metric will be able to demonstrate the equalities of schema document in terms of their corresponding attributes' ratios. Therefore the proposed metrics satisfy this property.

4) **Design Details are important (μ(A) ≡ μ(B) and μ(A) ≠ μ(B)):** This property means that the details of the structural design can change the metric value having the same functions. In **Figure 9**, A and B are the different schema documents having the same functionality, their attributes' values can be different because of different implementation. Therefore, Tr value is 0 because of the anonymous type in the schema A. In the schema B, Tr value is 1 because the type can be reused in the same schema or other schemas.

5) **Monotonic Complexity (μ (A) ≤ μ (AB) and μ (B) ≤ μ (AB)):** This means that the metric value of the combination schema can never be less than individual schema values. For example, ID_46 and ID 64 have the same EQ values and their combination schema value is equal individual schema values. The proposed metrics satisfy this property because the combination schema has the same value.



```
XMl schema document- A:
<element name="Movie">
 <complexType>
    <sequence>
<element name="Title"  type="xsd:string"/>
<element name="Director" type="xsd:string"/>
<element name="Genre"  type="xsd:string"/>
<element name="ReleaseYear"type="xsd:gYear"/>
    </ sequence>
  </ complexType>
 </ element>
```

```
XMl schema document- B:
 <complexType name="xsd:MovieType">
    <sequence>
<element name="Title"  type="xsd:string"/>
<element name="Director"  type="xsd:string"/>
<element name="Genre" type="xsd:string"/>
 <element name="ReleaseYear" type="xsd:gYear"/>
    </ sequence>
 </ complexType>
< element name="Movie" type="xsd: MovieType"/>
```

Figure 9. Listing of structural design A and design B.

6) **Non-equivalence of Interaction (μ(A) = μ(B) and μ (AC) = μ(BC)):** This states that even thought A and B has the same values, interaction between A and C and between B and C can be different resulting in different quality values. For example, ID_46 and ID_64 are the schema A and B. ID_60 lets the schema C and has different extensible quality in **Table 1**. The proposed metrics satisfy this property because the AC and BC schemas have different the EQ values with 20.3768 and 17.91042 respectively.

7) **Permutation Changes Complexity:** This property means that permutation impact on the metric value. The proposed metrics satisfy this property according to **Figure 9**.

8) **Renaming Property:** This property states that when the name of the measured schema component changes, the metric value can't be changed. The proposed metrics satisfy this property.

9) **Interaction Increases Complexity (μ(A) +μ(B) ≤ μ (AB)):** This property is that the interaction between two schemas cannot decrease their combination metric value. This property need not necessarily be true because of the percentage quality value of XML schema documents.

According to analytical analyzed result, the proposed metrics are robust with satisfying eight Weyuker's properties.

7. Conclusion

Three quality metrics: RQ, EQ and UQ are proposed to measure the reusable, extensible and understandable qualities of XML Schema documents based on binary entropy function and rank order centroid method respectively. The SOM method is used to cluster the XML schema documents to validate of the proposed three metrics empirically. To compare the matching clustering results of two types: depending on only one of each proposed metrics and depending on its based attributes are greater than 80%. According to the analyzed results empirically, the proposed metrics are more suitable to measure the three qualities than their based attributes. The metrics are evaluated by using Weyuker's 9 properties analytically. The RQ, EQ and UQ metrics are robust satisfying with eight properties. Therefore, the proposed metrics can provide valuable information for improving the quality of XML based system.

8. Acknowledgements

This work supported by University of Computer Studies, Mandalay, Myanmar. All teachers taught about this work and all teachers in my life are gracefully acknowledged.

REFERENCES

[1] T. M. Cover and J. A. Thomas, “Elements of Information Theory,” 2nd Edition, John Wiley & Sons, Inc., Hohoken, New Jersey and Published by Simultaneously in, 2006.

[2] F. Lourenco, V. Lobo and F. Bacao, “Binary Based Similarity Measures for Categorical Data and Their Applications in Self-Organizing Maps,” *JOCLAD* 2004-*XI Jornadas de Classificacao e Anlise de Dados*, Lisbon, 1-3 April 2004, pp. 1-8.

[3] E. J. Weyuker, “Evaluation Software Complexity Measures,” *IEEE Transactions on Software Engineering*, Vol. 14, No. 9, 1988, pp. 1357-1365. doi:10.1109/32.6178

[4] Y. Chen and R. McFadyen, “A DTD Complexity Metric,” *Proceedings of the* 21*st IASTED International Conference*, Applied Informatics, Innsbruck, 10-13 February 2001, Austria, pp. 1045-1052.

[5] S. Misra and B. Dilek, “Document Type Definition (DTD) Metrics,” *Romanian Journal of Information, Science and Technology*, Vol. 14, No. 1, 2011, pp. 31-50.

[6] M. Klettke, L. Schneider and A. Heuer, “Metrics for XML Document Collections,” *Lecture Notes in Computer Science*, Vol. 2490, 2002, pp. 15-28.

[7] A. McDowell, C. Schmidt and K. Bun Yue, “Analysis and Metrics of XML Schema,” *Proceedings of International Conference on Software Engineering Research and Practice*, 2004, pp. 538-544.

[8] D. Basci and S. Misra, “Measuring and Evaluating a Design Complexity Metric for XML Schema Documents,” *Journal of Information Science and Engineering*, Vol. 25, No. 5, 2009, pp. 1405-1425.

[9] D. Basci and S. Misra, “Entropy as a Measure of Quality of XML Schema Document,” *International Arab Journal of Information Technology*, Vol. 8, No. 1, 2010, pp. 75-83.

[10] D. Basci and S. Misra, “Complexity Metric for XML Schema Documents,” *Proceedings of the* 5*th International Workshop on SOA and Web Practices*, 2007, pp. 1-14.

[11] J. Visser, “Structure Metrics for XML Schema,” *Proceedings of XATA*, Portugal, 2006, pp. 1-12.

[12] B. Sumak, M. Hericko and M. Pusnik, “Towards a Framework for Quality XML Schema Evaluation,” *Proceedings of the ITI* 2007 29*th International Conference on Information Technology Interfaces*, Cavtat, 25-28 June 2007, pp. 783-788.

[13] D. Obasanjo, Microsoft Corporation, “W3C XML Schema Design Patterns: Avoiding Complexity,” Applies to: W3C XML Schema, 2003. http://www.xml.com

[14] H. S. Thompson, D. Beech, M. Maloney and N. Mendelsohn, “XML Schema Part 1: Structures Second Edition,” W3C Recommendation, 2004. http://www.w3.org/TR/xmlschema-1/

[15] D. Stephenson, “XML Schema Best Practice,” Hewlett-Packard Development Company, Palo Alto, 2004.

[16] “PESC Guidelines for XML Architecture and Data Modeling,” a Publication of the Postsecondary Electronic Standards Council (PESC), Version 3.0, 2005.

[17] T. Z. Thaw, “Measuring and Evaluation of Reusable Quality and Extensible Quality for XML Schema Documents,” 2011 *IEEE Student Conference on Research and Development*, Cyberjaya, 19-20 December 2011, pp. 473-478.

[18] D. Lee and W. W. Chu, “Comparative Analysis of Six XML Schema Languages,” *ACM SIGMOD Record*, Vol. 29, No. 3, 2000, pp. 76-87. doi:10.1145/362084.362140

Appendix

Table 1. The proposed metrics' values of XML schema document.

Document -ID	Schema Document Links	RQ	EQ	UQ
1	http://www.multispeak.org/interface/30j/10_OA_EA.asmx?WSDL	55.615	5.7096	0
2	http://services.nirvanix.com/ws/Accounting.asmx?WSDL	46.907	6.7668	0
3	http://services.argosoft.com/AddressValidation/AddressVerifier.asmx?WSDL	9.6491	0	0
4	https://api.channeladvisor.com/ChannelAdvisorAPI/v5/AdminService.asmx?WSDL	56.214	6.369	0
5	http://b3.caspio.com/ws/api.asmx?wsdl	65.717	21.981	0
6	http://www.oorsprong.org/websamples.arendsoog/ArendsoogbooksService.wso?WSDL	35.742	0	0
7	http://www.w3.org/Math/XMLSchema/mathml2/content/arith.xsd	68.793	0	7.5553
8	http://www.yazgelistir.com/YGServices/ArticleService.asmx?wsdl	28.313	1.79	0
9	http://omnovastage.crowechizekasp.com/attributes.asmx?wsdl	33.869	0	0
10	http://services.nirvanix.com/ws/Authentication.asmx?WSDL	8.6739	0	0
11	http://schemas.opengis.net/sensorML/1.0.1/base.xsd	15.195	7.6663	39.6267
12	http://banguat.gob.gt/variables/ws/*BDEF.asmx*?WSDL	46.649	0	0
13	http://www.webservicex.net/BibleWebservice.asmx?wsdl	10.309	0	0
14	http://www.thomas-bayer.com/axis2/services/BLZService?wsdl	60.155	0	0
15	http://www.mathertel.de/AJAXEngine/S02_AJAXCoreSamples/CalcService.asmx?WSDL	10.474	0	0
16	http://ww2.wso2.org/~charitha/calculator.xsd	9.6411	0	0
17	http://schemas.opengis.net/wms/1.3.0/*capabilities_1_3_0.xsd*	34.822	13.096	36.1851
18	http://ssl.9squared.com/catalog/catalog.asmx?WSDL	50.274	2.6522	0
19	http://webservice.webxml.com.cn/webservices/ChinaTVprogramWebService.asmx?WSDL	17.165	1.4874	0
20	http://service.ecocoma.com/convert/chinese.asmx?WSDL	24.61	0	0
21	http://service.ecocoma.com/geo/cityzip.asmx?WSDL	52.367	0	0
22	http://schemas.opengis.net/context/1.1.0/collection.xsd	55.225	26.21	0
23	http://svc.exaphoto.com/eXaPhoto/CollectionServices.asmx?WSDL	54.21	6.9871	0
24	http://www.w3.org/Math/XMLSchema/mathml2/presentation/common.xsd	57.318	36.255	55.5796
25	http://quiksilver.ws.eto.fr/Connexion.asmx?WSDL	9.1359	0	0
26	http://www.w3.org/Math/XMLSchema/mathml2/content/constants.xsd	68.707	0	10.069
27	http://www.esendex.com/secure/messenger/soap/ContactService.asmx?WSDL	37.636	4.5673	0
28	http://api.legiomedia.com/Content.asmx?WSDL	7.8981	2.0172	0
29	http://www.webservicex.net/ConverPower.asmx?WSDL	63.314	17.603	0
30	http://www.webservicex.net/ConvertTemperature.asmx?WSDL	63.314	17.603	0
31	http://www.webservicex.net/CovertPressure.asmx?WSDL	63.314	17.603	0
32	https://demo.docusign.net/API/3.0/Credential.asmx?WSDL	32.316	6.0195	0
33	http://www.webservicex.net/CreditCard.asmx?WSDL	12.259	0	0
34	http://ns.nsdl.org/schemas/MRingest/crsd_v1.06.xsd	77.124	26.416	70.7804
35	http://www.webservicex.com/CurrencyConvertor.asmx?wsdl	64.193	17.603	0

Continued

36	http://schemas.opengis.net/sld/1.1.0/*DescribeLayer.xsd*	72.356	54.989	17.8333
37	http://ws.interfax.net/dfs.asmx?WSDL	40.329	4.5992	0
38	http://service.ecocoma.com/geo/distance.asmx?WSDL	31.961	0	0
39	http://www.w3.org/Math/XMLSchema/mathml2/content/elementary-functions.xsd	67.026	0	6.0966
40	http://ws2.serviceobjects.net/ev/EmailValidate.asmx?WSDL	33.48	0	0
41	http://ws.cdyne.com/emailverify/Emailvernotestemail.asmx?wsdl	34.065	0	0
42	http://rangiroa.essi.fr:8080/dotnet/evaluation-cours/EvaluationWS.asmx?WSDL	26.43	0	0
43	http://schemas.opengis.net/wms/1.3.0/exceptions_1_3_0.xsd	75.543	21.969	0
44	http://schemas.opengis.net/filter/2.0/expr.xsd	66.985	13.249	25.4578
45	http://developer.factiva.com/2.0/wsdl/FDKParsers.wsdl	58.304	10.79	0
46	http://schemas.opengis.net/se/1.1.0/FeatureStyle.xsd	60.39	9.9698	25.4578
47	http://demo.soapam.com/services/FedEpayDirectory/FedEpayDirectoryService?WSDL	39.527	0	0
48	http://service.ecocoma.com/shipping/fedex.asmx?WSDL	46.052	0	0
49	http://schemas.opengis.net/filter/2.0/filterCapabilities.xsd	62.678	19.183	4.2523
50	http://www.w3.org/Math/XMLSchema/mathml2/content/functions.xsd	66.079	0	17.0165
51	http://schemas.opengis.net/sld/1.1.0/GetMap.xsd	65.385	40.3	10.8705
52	http://ns.nsdl.org/schemas/MRingest/harvest_v1.01.xsd.sav	43.316	27.14	71.0337
53	http://terraserver-usa.com/LandmarkService.asmx?WSDL	58.767	8.7403	0
54	http://www.w3.org/Math/XMLSchema/mathml2/content/linear-algebra.xsd	62.389	0	9.1801
55	http://www.w3.org/Math/XMLSchema/mathml2/content/logic.xsd	68.02	0	17.0165
56	http://www.chemspider.com/MassSpecAPI.asmx?WSDL	33.968	3.4028	0
57	http://www.w3.org/Math/XMLSchema/mathml3/mathml3-common.xsd	61.058	20.35	0
58	http://www.w3.org/Math/XMLSchema/mathml3/mathml3-strict-content.xsd	44.91	19.126	0
59	http://hooch.cis.gsu.edu/bgates/MathStuff/Mathservice.asmx?WSDL	8.9668	0	0
60	http://demo.turtletech.com/latest/webAPI/metering.asmx?WSDL	46.47	3.8511	0
61	http://schemas.opengis.net/sensorML/1.0.1/method.xsd	45.103	34.533	52.5938
62	http://www.exchangenetwork.net/repository/schema/NetDMR/1/0/NetDMR_Permits_v1.0.xsd	65.62	33.853	64.1722
63	http://ns.nsdl.org/schemas/nsdl_search/nsdl_search_v1.02.xsd	72.03	33.446	29.9817
64	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	56.068	9.9698	15.606
65	http://schemas.opengis.net/om/1.0.0/observation.xsd	66.783	22.001	83.316
66	https://www.devcallnow.com/WebService/OneCallNow.asmx?wsdl	28.053	5.6442	0
67	https://api.channeladvisor.com/ChannelAdvisorAPI/v3/OrderService.asmx?WSDL	70.996	18.636	0
68	http://schemas.opengis.net/ows/2.0/owsExceptionReport.xsd	63.52	39.154	88.2477
69	http://schemas.opengis.net/ows/2.0/owsGetCapabilities.xsd	66.939	32.962	70.9559
70	http://schemas.opengis.net/ows/2.0/owsGetResourceByID.xsd	65.537	37.598	83.582
71	http://trial.serviceobjects.com/pa/phoneappend.asmx?WSDL	45.482	0	0
72	http://service.thefamousgroup.com/ProjectService.asmx?wsdl	57.959	6.1383	0

Continued

73	http://ns.nsdl.org/schemas/provenance_about/provenance_about_v1.01.xsd	71.393	32.518	85.9227
74	http://www.partenairedejeu.fr/WebServices/RelationManager.asmx?WSDL	43.209	8.9402	0
75	http://www.hitslink.com/reportws.asmx?WSDL	66.304	14.151	0
76	http://ns.nsdl.org/schemas/ndr/response_v1.00.xsd	66.537	26.138	47.8354
77	http://www.oasis-open.org/committees/regrep/documents/2.0/schema/rs.xsd	30.297	18.692	34.95
78	http://www.w3.org/Math/XMLSchema/mathml2/presentation/scripts.xsd	62.389	0	9.1801
79	http://quisque.com/fr/chasses/blasons/search.asmx?WSDL	42.766	0	0
80	http://www.phdcc.com/findinsite/SearchService.asmx?wsdl	54.41	5.7096	0
81	http://www.w3.org/Math/XMLSchema/mathml2/content/semantics.xsd	64.339	2.1081	22.212
82	http://gw1.aql.com/soap/sendsmsservice.php?wsdl	67.379	18.206	0
83	https://portal.bmi.gv.at/ref/wsdl/zmr/test/wsdl/Service.wsdl	72.613	63.494	0
84	http://www.sipeaa.it/wset/ServiceET.asmx?WSDL	15.503	0	0
85	http://www.w3.org/Math/XMLSchema/mathml2/content/sets.xsd	68.004	6.0195	6.7416
86	http://www.geoservicios.com/V2.0/sgeo/sgeo.asmx?WSDL	44.88	0	0
87	http://e-commerce.pvc.maricopa.edu/cis234/old/ssilkey/SimpleAddress.xsd	60.447	24.868	0
88	http://www.w3.org/Math/XMLSchema/mathml2/presentation/space.xsd	68.355	22.673	36.7842
89	http://www.w3.org/Math/XMLSchema/mathml2/presentation/style.xsd	59.15	5.0705	41.689
90	http://msrmaps.com/TerraService2.asmx?WSDL	46.85	5.2641	0
91	http://www.w3.org/Math/XMLSchema/mathml2/content/tokens.xsd	76.467	17.603	15.606
92	http://www.webservicex.net/TranslateService.asmx?wsdl	63.715	17.603	0
93	http://www.w3.org/Math/XMLSchema/mathml2/content/vector-calculus.xsd	63.119	0	17.0165
94	http://services.test.musiccue.net/rapidcueapplication/WorkManager.asmx?WSDL	9.7771	0	0
95	http://www.imagine-r.com/services/WsImagineR.asmx?WSDL	46.628	2.7369	0
96	http://www.xignite.com/xMetals.asmx?WSDL	58.867	16.564	0
97	http://www.xignite.com/xNASDAQLastSale.asmx?WSDL	62.755	18.449	0
98	http://www.xignite.com/xNews.asmx?WSDL	56.067	11.206	0
99	http://www.xignite.com/xQuotes.asmx?WSDL	58.823	14.558	0
100	http://www.xignite.com/xwatchlists.asmx?WSDL	61.415	16.265	0