Abstract—Web Services is a technology for building distributed software applications that are built upon a set of information and communication standards. Among those standards is the Web Services Description Language (WSDL) which is an XML-based language for describing service descriptions. Service providers will publish WSDL documents of their Web services so that service consumers can learn about service capability and how to interface with the services. Since WSDL documents are the primary source of service information, readability of WSDL documents is of concern to service providers, i.e. service descriptions should be understood with ease by service consumers. Providing highly readable service descriptions can then be used as a strategy to attract service consumers. However, given highly readable information in the WSDL documents, competitors are able to learn know-how and can copy the design to offer competing services. Security attacks such as information espionage, client impersonation, command injection, and denial of service are also possible since attackers can learn about exchanged data and invocation patterns from WSDL documents. While readability of service descriptions makes Web services discoverable, it contributes to service vulnerability too. Service designers therefore should consider this trade-off when designing service descriptions. Currently there is no readability measurement for WSDL documents. We propose an approach to such measurement so that service designers can determine if readability is too low or too high with regard to service discoverability, service imutation, and service attack issues, and then can consider increasing or lowering service description readability accordingly. Our readability measurement is based on the concepts or terms in service domain knowledge. Given a WSDL document as a service description, readability is defined in terms of the use of difficult words in the description and the use of words that are key concepts in the service domain. As an example, we measure readability of the WSDL document of public Web services, and outline a method to lower or increase readability.

Keywords— Concept Hierarchy, Ontology, Readability, Web Services, WSDL.

I. INTRODUCTION

Web Services is a technology for building distributed software applications. The building blocks are services which are software units that are built upon a set of information and communication standards. Among those standards is the Web Services Description Language (WSDL) which is an XML-based language for describing service descriptions. A WSDL document is defined by a service provider and used by service consumers in discovering service capability and establishing interaction between consumer-side applications and the Web service. The structure of a WSDL document (version 1.1) which describes what the service is capable of and what data are exchanged comprises the XML elements <types>, <message>, <portType>, and <documentation> [1].

Since WSDL documents are the primary source of service information, readability [2] of WSDL documents or Web services descriptions is of concern to service providers, i.e. service descriptions should be understood with ease by service consumers. Meaningful names should be given to the service interface, operations, input and output messages, and data. In addition, sufficient documentation should be provided regarding functional scope and limitation of use. Providing well-defined readable service descriptions can be used as a strategy by service providing organizations to attract service consumers.

Despite being desirable, readability of service descriptions has its downside. Other organizations can gather information from a service WSDL document to learn know-how and then copy the design to offer competing services. Also, published WSDL documents can provide security attackers with information like schemas of exchanged data, invocation patterns, and service location. Attackers may be able to guess other private operations. This leads to more serious attacks such as information espionage, client impersonation, command injection, and denial of service.

While readability of service descriptions makes Web services discoverable, it contributes to service vulnerability too. Service designers therefore should consider this trade-off when designing service descriptions. Currently there is no readability measurement for WSDL documents. We therefore aim to propose an approach to such measurement so that service designers can determine if readability is too low or too high with regard to service discoverability, service imitation, and service attack issues, and then can consider increasing or lowering service description readability accordingly.

We apply the concept-based readability measurement model proposed by Yan et al. [3] to the context of Web service descriptions. As the name implies, our readability measurement is based on the concepts or terms in service domain knowledge. Given a WSDL document as a service description, readability is defined in terms of the use of difficult words in the description and the use of words that are
key concepts in the service domain. For example, if the service description contains simple words or closely-related terms within the domain, it should be easy to understand the functionality of the service from the service description. Here, service domain knowledge is described as ontology [4] that defines vocabulary of concepts and properties as well as their relationships, using an XML-based OWL language [5]. Readability assessment can be conducted by the quality assurance team or service designers who have knowledge of the service domain, and can be of several uses. The assessors can compare readability of their WSDL documents with that of the competing services and may evaluate if readability should be improved to attract more consumers. In a certain case, the assessors may consider adjusting the service descriptions if security issues are of concern. We also outline a method to increase and lower readability of service descriptions.

The rest of this paper is organized as follows. Section II discusses related work and Section III presents the concept-based document readability model. We propose the methodology to assess readability of WSDL documents in Section IV and applies the methodology to WSDL documents of five Web services in the e-commerce domain for comparison in Section V. Section VI concludes the paper with a discussion of the approach and future outlook.

II. RELATED WORK

By definition, readability means “the level of ease or difficulty with which text material can be understood by a particular reader who is reading that text for a specific purpose” [2]. Readability is dependent upon many characteristics of both the text and the readers, and its concept has been applied to many kinds of text material including books, technical documents, online documents, and Web pages. Many formulas for measuring text readability are available and most of them deal with only text features. That is, texts that use difficult words are more difficult to understand than those with simple words, and texts with long sentences and complex syntax are difficult to read.

Yan et al. [3] propose a different but interesting approach to measuring text readability in the context of online documents. They argue that not only domain experts but also average users are searching more and more for domain-specific information from online documents, particularly in the medical area, and these documents are of different readability level. However, traditional readability formulas are designed for general purpose texts and insufficient to deal with technical materials in a specific domain. Therefore, for the document ranking purpose, their model takes advantages of a traditional readability formula and domain knowledge to measure readability of domain-specific documents at the word level, i.e. it focuses on how the domain-specific terms in a document affect readability of the document. For example, if the document contains closely-related terms in the domain vocabulary, it should be more easily readable and comprehensible to readers of that domain. In their approach, the domain knowledge is represented as a concept hierarchy or ontology.

Zhao and Kan [6] argue that the ontology-based approach such as [3] has a limitation in that it requires expert knowledge which is still expensive and not readily available in most domains. They hence present an iterative computation on the WSDL document based on the intuition that both the text and the readers, and its concept has been applied to many kinds of text material including books, technical documents, online documents, and Web pages. Many formulas for measuring text readability are available and most of them deal with only text features. That is, texts that use difficult words are more difficult to understand than those with simple words, and texts with long sentences and complex syntax are difficult to read.

Yan et al. [3] propose a different but interesting approach to measuring text readability in the context of online documents. They argue that not only domain experts but also average users are searching more and more for domain-specific information from online documents, particularly in the medical area, and these documents are of different readability level. However, traditional readability formulas are designed for general purpose texts and insufficient to deal with technical materials in a specific domain. Therefore, for the document ranking purpose, their model takes advantages of a traditional readability formula and domain knowledge to measure readability of domain-specific documents at the word level, i.e. it focuses on how the domain-specific terms in a document affect readability of the document. For example, if the document contains closely-related terms in the domain vocabulary, it should be more easily readable and comprehensible to readers of that domain. In their approach, the domain knowledge is represented as a concept hierarchy or ontology.

Zhao and Kan [6] argue that the ontology-based approach such as [3] has a limitation in that it requires expert knowledge which is still expensive and not readily available in most domains. They hence present an iterative computation on the WSDL document based on the intuition that both the text and the readers, and its concept has been applied to many kinds of text material including books, technical documents, online documents, and Web pages. Many formulas for measuring text readability are available and most of them deal with only text features. That is, texts that use difficult words are more difficult to understand than those with simple words, and texts with long sentences and complex syntax are difficult to read.

Yan et al. [3] propose a different but interesting approach to measuring text readability in the context of online documents. They argue that not only domain experts but also average users are searching more and more for domain-specific information from online documents, particularly in the medical area, and these documents are of different readability level. However, traditional readability formulas are designed for general purpose texts and insufficient to deal with technical materials in a specific domain. Therefore, for the document ranking purpose, their model takes advantages of a traditional readability formula and domain knowledge to measure readability of domain-specific documents at the word level, i.e. it focuses on how the domain-specific terms in a document affect readability of the document. For example, if the document contains closely-related terms in the domain vocabulary, it should be more easily readable and comprehensible to readers of that domain. In their approach, the domain knowledge is represented as a concept hierarchy or ontology.

Zhao and Kan [6] argue that the ontology-based approach such as [3] has a limitation in that it requires expert knowledge which is still expensive and not readily available in most domains. They hence present an iterative computation on the WSDL document based on the intuition that both the text and the readers, and its concept has been applied to many kinds of text material including books, technical documents, online documents, and Web pages. Many formulas for measuring text readability are available and most of them deal with only text features. That is, texts that use difficult words are more difficult to understand than those with simple words, and texts with long sentences and complex syntax are difficult to read.

Yan et al. [3] propose a different but interesting approach to measuring text readability in the context of online documents. They argue that not only domain experts but also average users are searching more and more for domain-specific information from online documents, particularly in the medical area, and these documents are of different readability level. However, traditional readability formulas are designed for general purpose texts and insufficient to deal with technical materials in a specific domain. Therefore, for the document ranking purpose, their model takes advantages of a traditional readability formula and domain knowledge to measure readability of domain-specific documents at the word level, i.e. it focuses on how the domain-specific terms in a document affect readability of the document. For example, if the document contains closely-related terms in the domain vocabulary, it should be more easily readable and comprehensible to readers of that domain. In their approach, the domain knowledge is represented as a concept hierarchy or ontology.
A well-known traditional readability formula is the Dale-Chall’s readability index [10]. This index sees that, the length of the sentences in a document and the difficulty of words correlate with the difficulty of reading material. Since the concept-based readability model measures readability at the word level, sentence-level complexity is not applicable and hence only word difficulty is considered. To determine the difficulty of words, words in the document are identified as either familiar or unfamiliar words. That is, they are familiar words if they can be found in the Dale list of approximately 3,000 familiar words. Otherwise, they are unfamiliar, and hence difficult words. The simplified version of the Dale-Chall’s readability index of a WSDL document \( d_i \) can be computed by (5):}

\[
DaCw(d_i) = PDW
\]  

where \( PDW \) = percentage of difficult words in the WSDL document \( d_i \) (i.e. number of difficult words divided by number of words and multiplied by 100).

### IV. CONCEPT-BASED READABILITY ASSESSMENT METHODOLOGY

The concept-based readability of a WSDL document can be measured by an assessor who is a member of the quality assurance team or a service designer. The assessor must have knowledge of the service domain in order to choose (or construct) the concept hierarchy or domain ontology appropriately, and at the end evaluate the readability score. The assessment methodology comprises the following steps.

#### A. Service Information Preparation

To prepare for assessment, the assessor first does the following.

1. **Select WSDL Document**: The assessor selects a Web service and acquires its WSDL document. Note that if the assessor wants to compare readability of two Web services, both services will be assessed individually but they must be in the same domain and share the same concept hierarchy.

2. **Obtain Concept Hierarchy of Service Domain from Ontology Library**: The concept hierarchy can be a domain ontology defined by domain experts. We consult either search engines or Web sites that publish domain-specific ontologies, e.g. [11]-[13], to discover relevant ontology for the Web service. In the case that multiple ontologies of the service domain are found, a tool like Protégé [14] can be used by the assessor to merge them into a single integrated ontology.

#### B. Readability Measurement

Generally the assessor measures readability of the whole WSDL document, but in some case it might be useful as well to measure readability of particular WSDL elements for a more detailed analysis. The assessor can first select to measure readability of either the whole document or certain elements, and then apply the concept-based readability model (6). In the case that the measurement targets a particular

---

**A. Document Scope (DS)**

Document scope is defined as the coverage of the domain concepts in the document. The coverage is viewed from two angles. First, the more the document contains domain terms, the less readable the document tends to be since the document is likely to contain a larger number of specific concepts. Second, the deeper the domain terms appear in the concept hierarchy, the more difficult the document is to read. The document scope of a WSDL document \( d_i \) can be computed by (1):

\[
Scope(d_i) = e^{- \sum_{i=1}^{n} \text{depth}(c_i)}
\]

where \( \text{depth}(c_i) = \text{depth of domain concept } c_i \) in the WSDL document \( d_i \), with regard to the concept hierarchy.

**B. Document Cohesion (DC)**

Document cohesion refers to how focused the text is on a particular topic. It can be computed by the semantic relatedness between the domain terms in the document which is reflected by the links (or shortest path) between them with respect to the given concept hierarchy. The more cohesive the domain terms in the document are, the more readable the document is. The document cohesion of a WSDL document \( d_i \) can be computed by (2)-(4):

\[
\text{Cohesion}(d_i) = \frac{\sum_{i=1,j=1}^{n} \text{Sim}(c_i,c_j)}{\text{Number of Associations}}, \text{ where } n > 1, i < j
\]

\[
\text{Sim}(c_i,c_j) = - \log \frac{\text{len}(c_i,c_j)}{2D}
\]

\[
\text{Number of Associations} = \frac{n(n-1)}{2}
\]

where \( \text{len}(c_i,c_j) = \text{shortest path between } c_i \text{ and } c_j \) in the concept hierarchy, \( D = \text{maximum tree depth in the concept hierarchy} \), and \( n = \text{total number of domain concepts in the WSDL document } d_i \).

**C. Simplified Dale-Chall’s Readability Index (DaCw)**

A well-known traditional readability formula is the Dale-Chall’s readability index [10]. This index sees that, the length of the sentences in a document and the difficulty of words correlate with the difficulty of reading material. Since the concept-based readability model measures readability at the word level, sentence-level complexity is not applicable and hence only word difficulty is considered. To determine the difficulty of words, words in the document are identified as either familiar or unfamiliar words. That is, they are familiar words if they can be found in the Dale list of approximately 3,000 familiar words. Otherwise, they are unfamiliar, and hence difficult words. The simplified version of the Dale-Chall’s readability index of a WSDL document \( d_i \) can be computed by (5):

\[
DaCw(d_i) = PDW
\]
WSDL element, the content of the element corresponds to a textual document of the model.

**C. Evaluation**

Once the readability score is obtained, the assessor evaluates if readability of the whole WSDL document (or a particular WSDL element) is appropriate. The assessor may be concerned about attracting the consumers and at the same time being vulnerable to attackers and competitors. The assessor can use the readability score for comparison purpose, e.g. comparing with the scores of competitors’ services. The scores can be adjusted if the assessor sees fit.

**D. Readability Adjustment**

The concept-based readability score is dependent on the design and naming of terms within the WSDL document and the quality of the concept hierarchy. That is, the assessor should choose service domain ontology from a reliable source. Measuring readability of specific WSDL elements can help the assessor to pinpoint which parts of the WSDL document should be redesigned and how to adjust their contents in order to increase or lower the score appropriately. According to (6), it is apparent that we can adjust the readability score by redesign or renaming of terms within the WSDL document since change of terms can affect the document scope, document cohesion, and simplified Dale-Chall’s readability index in the following ways.

- **Domain Scope**: For a WSDL document, adding terms or changing to deeper terms in the concept hierarchy will decrease its readability score.
- **Domain Cohesion**: For a WSDL document, adding terms or changing to terms that are more closely associated with respect to the concept hierarchy will decrease the shortest path and hence increase its readability score.
- **Simplified Dale-Chall’s readability index**: For a WSDL document, adding or changing to terms in the Dale list of familiar words will increase its readability score.

Therefore, to increase the readability score to improve service discoverability, the assessor may, where appropriate,

- Change domain terms in the WSDL document to non-domain terms;
- Change domain terms in the WSDL document to those that appear shallower in the concept hierarchy;
- Change or add terms so that the WSDL document contains domain terms that are more closely associated by semantics with respect to the concept hierarchy;
- Change or add terms so that the WSDL document contains more of Dale’s familiar words.

On the other hand, to lower the readability score to make the service description less comprehensible to competitors and attackers, the assessor may, where appropriate,

- Change non-domain terms in the WSDL document to domain terms;
- Change domain terms in the WSDL document to those that appear deeper in the concept hierarchy;
- Change or add terms so that the WSDL document contains domain terms that are more loosely associated by semantics with respect to the concept hierarchy;
- Change or add terms so that the WSDL document contains less of Dale’s familiar words.

After redesign or renaming of terms within the WSDL document, the assessor will repeat the measurement process to obtain the adjusted score. The adjustment and measurement can be repeated as necessary until the assessor is satisfied with the score.

**V. APPLICATION OF READABILITY ASSESSMENT METHODOLOGY**

The methodology in Section IV is applied to assess readability of WSDL documents of five service providers in the e-commerce domain. They are Amazon [15], ClickandBuy [16], eBay [17], KonaKart Enterprise [18], and PayPal [19]. The e-commerce ontology from GoodRelations [20], used by over 10,000 small and large shops worldwide, is selected to represent the concept hierarchy for the assessment.

**A. Readability Assessment Walk-Through**

Due to space limitation, we present an example of how to apply the concept-based readability model to Amazon’s WSDL document [15]. The following steps are performed by our WSDL readability calculator tool (see Section V-C).

Step 1: The tool will process the WSDL document by extracting WSDL element names; WSDL syntax (such as tag names, types, cardinality) is excluded. Duplicate terms are also removed. In total, there are 379 extracted terms, and Table 1 shows some of the terms that are extracted from a WSDL snippet in Figure 1.

![Figure 1. Snippet of Amazon’s WSDL](image)

**Table 1. Example of Extracted Terms from Amazon’s WSDL**

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PurchaseURL</td>
<td>SimilarProducts</td>
</tr>
<tr>
<td>MobileCartURL</td>
<td>TopSellers</td>
</tr>
<tr>
<td>SubTotal</td>
<td>TradeInValue</td>
</tr>
<tr>
<td>CartItems</td>
<td>SimilarViewedProducts</td>
</tr>
<tr>
<td>SavedForLaterItems</td>
<td>LowestNewPrice</td>
</tr>
</tbody>
</table>

![xml sequence](xml)

Step 2: From the list of extracted terms, the tool will process each term to further extract key individual words, change them all into a singular form, and remove duplicate words and any single character. For example, the term
“CartItems” will be extracted into two individual words “Cart” and “Item”. In total, there are 280 individual words, and Table 2 shows some of them.

<table>
<thead>
<tr>
<th>Purchase</th>
<th>Url</th>
<th>Mobile</th>
<th>Cart</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub</td>
<td>Total</td>
<td>New</td>
<td>Price</td>
</tr>
<tr>
<td>Saved</td>
<td>For</td>
<td>Later</td>
<td>Viewed</td>
</tr>
<tr>
<td>Similar</td>
<td>Product</td>
<td>Top</td>
<td>Seller</td>
</tr>
<tr>
<td>Item</td>
<td>Trade</td>
<td>In</td>
<td>Value</td>
</tr>
</tbody>
</table>

Step 3: Individual words will be mapped to the concepts in the GoodRelations concept hierarchy [21]. In total, there are 16 individual words that can find a match in the concept hierarchy, and Table 3 shows some of them.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td>3</td>
</tr>
<tr>
<td>Product</td>
<td>2</td>
</tr>
<tr>
<td>Value</td>
<td>1</td>
</tr>
<tr>
<td>Price</td>
<td>1</td>
</tr>
</tbody>
</table>

Step 4: To calculate the document scope, the depth of each domain concept found in the WSDL document is determined with regard to the concept hierarchy. Table 4 shows the depth of some domain concepts. Given (1) in Section III, the document scope of Amazon’s WSDL document is 2.789468092868928E-10.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td>3</td>
</tr>
<tr>
<td>Product</td>
<td>2</td>
</tr>
<tr>
<td>Value</td>
<td>1</td>
</tr>
<tr>
<td>Price</td>
<td>1</td>
</tr>
</tbody>
</table>

Step 5: To calculate the document cohesion, the shortest path between each pair of all 16 domain concepts that are found in the WSDL document is determined, with respect to the concept hierarchy of depth 5. Table 5 shows the shortest path between some pairs of domain concepts. Given (2)-(4) in Section III, the document cohesion of Amazon’s WSDL document is 0.22815030517112506.

<table>
<thead>
<tr>
<th>c_i</th>
<th>c_j</th>
<th>len(c_i, c_j)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td>Product</td>
<td>6</td>
</tr>
<tr>
<td>Item</td>
<td>Value</td>
<td>5</td>
</tr>
<tr>
<td>Item</td>
<td>Price</td>
<td>5</td>
</tr>
<tr>
<td>Product</td>
<td>Value</td>
<td>5</td>
</tr>
<tr>
<td>Product</td>
<td>Price</td>
<td>4</td>
</tr>
</tbody>
</table>

Step 6: To calculate the simplified Dale-Chall’s readability index, there are 87 out of 280 individual words that are found in the Dale list of familiar words; Table 6 shows some of them. Therefore there are 193 words that are considered as difficult words. Given (5) in Section III, the simplified Dale-Chall’s readability index of Amazon’s WSDL document is 68.92857142857143.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td>3</td>
</tr>
<tr>
<td>Product</td>
<td>2</td>
</tr>
<tr>
<td>Value</td>
<td>1</td>
</tr>
<tr>
<td>Price</td>
<td>1</td>
</tr>
</tbody>
</table>

Step 7: Finally, given (6) in Section III, the tool will calculate the concept-based readability score of Amazon’s WSDL document. The score is 0.24265807747079726.

B. Readability Assessment Result

Readability scores of the whole WSDL documents of five service providers are shown in Figure 2. PayPal’s Web Service provides the most readable WSDL document whereas readability of ClickandBuy’s WSDL is the lowest. Readable and comprehensible service description can be one factor in the popularity of PayPal Web service, while WSDL readability of KonaKart Enterprise, Amazon, and eBay follows PayPal’s very closely, indicating that they can be competing Web services providing similarly readable service descriptions. On the other hand, ClickandBuy’s WSDL document is likely to expose less information to competitors and attackers.

![Figure 2. Readability scores of five service WSDLs](image)

C. WSDL Readability Calculator

The readability assessment above is supported by a tool called WSDL Readability Calculator as shown in Figure 3. The tool is developed by using Eclipse Java EE IDE [21] and the Protégé Java library [14] is used to read the domain ontology file.

The tool requires the following input from the assessor:
- Web service description URL or a WSDL file;
- Service domain ontology URL or an OWL file;
- Calculation method, i.e. whole document or specific WSDL element.

Once the assessor inputs all service details and starts the calculation, the tool will extract individual words from the WSDL document and then extracts concepts from the domain ontology before calculating the document scope, document cohesion, and simplified Dale-Chall’s readability index for the
Our future work is to experiment with a number of service providing organizations in Thailand to see if the approach can help them with the design of service descriptions and how they use the method that is outlined in this paper to adjust readability of their WSDL documents.

REFERENCES


